Advances in Statistical and Machine Learning Methods for Image Data, with Application to Alzheimer's Disease

by

Tianyu Ding

BS, Applied Mathematics, South China University of Technology, China, 2013

MPH, Biostatistics, Emory University, 2015

Submitted to the Graduate Faculty of
the Graduate School of Public Health in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF BIOSTATISTICS

This dissertation was presented

by

Tianyu Ding

It was defended on

May 20, 2020

and approved by

Robert Krafty, PhD, Associate Professor, Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dana Tudorascu, PhD, Associate Professor, Department of Psychiatry and Biostatistics, School of Medicine, University of Pittsburgh

Ann Cohen, PhD, Associate Professor, Department of Psychiatry, School of Medicine, University of Pittsburgh

Stewart Anderson, PhD, Professor, Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Advisors: Robert Krafty, PhD, Associate Professor, Biostatistics, Graduate School of Public Health, University of Pittsburgh,

Dana Tudorascu, PhD, Associate Professor, Department of Psychiatry and Biostatistics, School of Medicine, University of Pittsburgh

Advances in Statistical and Machine Learning Methods for Image Data, with Application to Alzheimer's Disease

Tianyu Ding, PhD

University of Pittsburgh, 2020

Abstract

The revolutionary development of neuroimage technology allows for the generation of large-scale neuroimage data in modern medical studies. For example, structural magnetic resonance imaging (sMRI) is widely used in segmenting neurodegenerative regions in the brain and positron-emission tomography (PET) is commonly used by clinicians and researchers to quantify the severity of Alzheimer's disease.

In the first part of this dissertation, we build "OASIS-AD", which is a supervised learning model based on a well-validated automated segmentation tool "OASIS" in multiple sclerosis (MS). OASIS-AD considers the specific challenges raised by WMH in Alzheimer's Disease (AD) to reduce false discoveries. We show that OASIS-AD performs better than several existing automated white matter hyperintensity segmentation approaches.

In the second part of this dissertation, we develop an interpretable penalized multivariate high-dimensional method for image-on-scalar regression that can be used for association studies between high-dimensional PET images and patients' scalar measures. This method overcomes the lack of interpretability in regularized regression after reduced-rank decomposition through a novel encoder-decoder based penalty to regularize interpretable image characteristics. Empirical properties of the proposed approach are examined and compared to existing methods in simulation studies and in the analysis of PET images from subjects in a study of Alzheimer's Disease.

In the third part of this dissertation, we developed ACU-Net, an efficient convolutional network for medical image segmentation. The proposed deep learning network overcomes the small sample size problem of training a deep neural network when used for medical image segmentation. It also decreases computation cost by increasing the effective degrees

of freedom through data augmentation and the novel use of convolutional layers blocks to compress the model. We show that ACU-Net can achieve competitive performance while dramatically decreases the computation cost compared with modern CNNs.

Public health significance: This dissertation proposes new statistical and machine learning methods for two aging-related problems: (1) automatically segmenting white matter hyperintensity (WMH), a biomarker of neurodegenerative pathology, and (2) estimating the association between neurodegeneration pathology and vascular measures, which are important to aging population living quality and can be studied by clinical neuroimage data.

Table of Contents

List of Tables

List of Figures

Preface

This basis for this research originally stemmed from my passion for developing better methods for studying Alzheimer's disease structure and anatomy with neuroimages and for understanding their association with other types of data. As the world moves further into the digital age, generating vast amounts of neuroimages and other large datasets, there will be more but complicated resources that researchers can access. How can we utilize those resources to build statistical models to help reveal the pathology of brain diseases? It is my passion to not only find out but to develop tools to help researchers with both statistical and biological interpretation.

In truth, I could not have achieved my current level of success without a strong support group. First of all, my parents, who supported me with love and understanding. Secondly, my girlfriend, who is always with me no matter what situations we come across. Lastly, my committee members, each of whom has provided patient advice and guidance throughout the research process.

1.0 Introduction

The recent explosion in the number of studies that collect neuroimage data has led to an increased need in statistical models and methods for their analysis. Among these studies, two main directions are (1) segmentation between normal and abnormal regions based on different modalities of neuroimage data (Caligiuri et al., 2015), and (2) association studies and predictive modeling between medical image data and other types of data, such as demographic and genetic data (Bigos and Weinberger, 2010).

The first project of this dissertation focuses on segmentation between white matter hyperintensities (WMHs) and normal brain tissue based on a normal aging cohort (Nadkarni et al., 2019). WMHs are areas in the white matter of the brain that appear hyperintense on a T2-weighted-Fluid-Attenuated Inversion Recovery (T2-FLAIR) scan and appear hypointense on a T1-weighted scan as compared to normal appearing white matter. Existence of WMHs can be very challenging when using traditional automatic MRI processing techniques for brain images of older adults. For example, segmentation of brain imaging data into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) is a crucial processing step in brain imaging studies. Existing automatic segmentation approaches were developed for the brain images of healthy young adults, which generally do not contain WMHs. When WMHs are present, automatic segmentation tools that use T1-weighted images may incorrectly classify WMHs as GM since both appear hypointense. Therefore, large WMH areas could artificially increase the estimated GM volumes in cross-sectional studies and could lead to severe underestimation of GM loss in longitudinal studies. This can be a major problem in studies that use GM volume as a general marker of brain atrophy. Moreover, WMHs are more prevalent in older adults and women (Van Den Heuvel et al., 2004), which may lead to differential tissue classification performance in specific subgroups.

WMHs appear in a variety of studies, both in individuals who are clinically symptomatic or asymptomatic. In particular, WMHs are pervasive in studies of aging, Alzheimer's Disease (AD), bipolar disorder (Pillai et al., 2002), and stroke (Wong et al., 2002). WMH segmentation is crucial for correcting tissue classification as well as for estimating the WMH volume

directly, as this is often used as a marker of cerebrovascular diseases. In this dissertation we focus on WMH segmentation in the aging brain in general and aging brain affected by AD in particular.

Thus, we propose OASIS-AD, an automatic supervised approach based on logistic regression and careful consideration of brain spatial information. OASIS-AD is an approach evolved from OASIS (Sweeney et al., 2013)(Automated Statistical Inference for Segmentation), which was developed for automatic lesion segmentation in multiple sclerosis (MS). OASIS-AD is a major refinement of OASIS that takes into account the specific challenges raised by WMH, in particular, in AD. A common problem in WMHs segmentation tool is false-positives. In the original OASIS, voxels are selected naively in prepocessing steps through the top 15% FLAIR intensities, which might be appropriate in MS, but not in AD. OASIS-AD changes the image preprocessing steps and adds three novel processing steps to reduce false-positives.

The second project introduces a novel interpretable regularized image-on-scalar regression within a reduced-rank regression framework, which can be used in both association studies and predictive modeling between high-dimensional neuroimaging data and scalar data. Compared to scalar-on-image regression, image-on-scalar regression uses scalar data to predict image data. As images are often more difficult to obtain than scalar values, it provides a means of conducting inference on phenomena that are usually quantified through costly image data with more readability available data. For example, our motivating study of Alzheimer's Disease (Cohen et al., 2013) is concerned with understanding connections between positron-emission tomography (PET) images, which are used by clinicians and researchers to quantify anatomical symptoms of Alzheimer's disease, with easily obtainable correlates of dementia, such as psychosocial measures and blood pressure. Image-on-scalar regression is particularly challenging since it uses low-dimensional data to predict high-dimensional data, and since associations are often sparse with weak signals at a set of particular voxels.

In this project, we propose interpretable reduced-rank regression (IRRR) as a method for image-on-scalar regression. The method uses a fused sparse group lasso penalty after dimension reduction, which reduces the size of the high-dimensional model while regularizing based on spatial smoothness, structural and functional grouping, and sparsity. The penalty

includes an encoder-decoder to enable it to be formulated on the reduced-rank space, but maintain biological interpretation and regularize on the image space.

The third project introduces a compact deep neural network architecture. Deep learning architectures have recently achieved great success on problems in nature language processing and computer vision. Among these, convolutional neural network (CNN), which are generally built with convolutional layers, pooling layers and fully-connected layers (O'Shea and Nash, 2015), are widely used in image classification and segmentation. However, as described in Miotto et al. (2017), medical data such as imaging, genetics, and electronic health records are complex, heterogeneous, poorly annotated and generally unstructured. This commonly leads to complicated data with lack of sufficient domain knowledge when directly applying end-to-end deep learning models. The goal of this project is to mitigate these issues when applying modern deep learning architectures to biomedical image segmentation, such as WMH segmentation, by overcoming two common challenges. The first obstacle is the high resolutions but low sample sizes faced with general image classification or segmentation problems (Deng et al., 2009). The second is heavy computation cost for a well-trained deep neural network. Our goal is to build a scalable state-of-the-art deep learning model for medical image studies.

In this project, we develop a novel compressed convolutional neural network architecture based on U-Net (Ronneberger et al., 2015). U-Net is a well-validated biomedical image segmentation network which utilizes a symmetric auto-encoder architecture and data augmentation to increase efficiency with small samples. As a well-known property, successful training of deep networks requires thousands of well-labeled training samples, which are usually unavailable in medical image areas, especially for sMRI. The data augmentation used in U-Net partially lowers the number of required image samples to train a reliable network. In addition, inspired by our second project, we incorporated multiple modern techniques related to dimension reduction and decomposition to build an asymmetric auto-encoder to decrease computation cost while remain the competitive accuracy compared with original neural network architectures.

2.0 White Matter Hyperintensity Detection in Alzheimer's Disease

2.1 Introduction

As discussed in the previous chapter, WMH segmentation is essential in the analysis of neuroimage data of elderly subjects. A review of existing WMH segmentation methods is provided in Caligiuri et al. (2015). The methods can be divided into three categories: (1) supervised learning algorithms using manually-labeled tracings of WMHs, (2) unsupervised learning algorithms using unlabeled manual tracings, and (3) semi-automated algorithms with various degrees of user intervention. Supervised classification algorithms include: k-nearest neighbors (kNN), non-parametric classification using the k closest training samples in the feature space (Anbeek et al., 2004), support vector machines (SVM) (Lao et al., 2008), Bayesian methods that combine multivariate signal intensity and spatial information (Herskovits et al., 2008), artificial neural networks (ANN) using multi-sequence images (Dyrby et al., 2008), Gaussian mixture models (Simões et al., 2013), logistic regression of multi-sequence images (Sweeney et al., 2013), adaptive intensity threshold search (Yoo et al., 2014), and deep convolutional neural networks (Ghafoorian et al., 2017). Unsupervised classification algorithms include: a two-level fuzzy inference system based on proton density (PD) and T2-FLAIR images (Admiraal-Behloul et al., 2005), a fuzzy connected algorithm combined with image registration (Wu et al., 2006), and a geostatistical fuzzy c-means clustering algorithm (Anitha et al., 2012). Semi-automated algorithms include: region growing using adaptive thresholding (Itti et al., 2001), bispectral fuzzy class means (Sheline et al., 2008), and semi-automatic peak identification on the 2D histogram of T1 and T2 intensities (Sheline et al., 2008). Caligiuri et al. (2015) concluded that a good WMH segmentation method should include a comprehensive image preprocessing pipeline based on multi-sequence data that takes into account spatial information about lesions and corrects for false positives.

In real application, T2-FLAIR images are produced by using very long TE and TR times, where repetition time (TR) is the amount of time between successive pulse sequences applied to the same slice and time to echo (TE) is the time between the delivery of the RF pulse

and the receipt of the echo signal. This sequence is very sensitive to pathology and makes the differentiation between CSF and an abnormality much easier. Since WMHs are bright on T2-FLAIR, researchers use this modality to do manually segmentation for more accurate segmentation and better visualization purposes.

Our proposed method "OASIS-AD", an automatic supervised approach evolved from OASIS (Sweeney et al., 2013), is developed by incorporating three novelties on both data processing step and modeling steps to increase classification accuracy of WMHs and to reduce false-positives. First, it uses an eroding procedure on the skull stripped mask, which can remove small spurious bright spots (salt noise) in images. Second, it incorporates an nearest neighbor feature construction approach, which utilizes the spacious information of a 3D brain image to refine segmentation probability map to reduce false positives. Lastly, it uses a Gaussian filter to smooth segmentation probability map to reduce false positives. We show that OASIS-AD performs better than existing WMH segmentation approaches when compared to manually segmentation by our experienced radiologists, the generally accepted gold standard.

2.2 Materials and Methods

In this section, we introduce the steps of OASIS-AD with details. OASIS-AD has three main components: (1) development of a binary brain tissue mask, (2) normalization of MRI intensities and creation of smoothed volumes, and (3) two-step modeling. The first step of modeling consists of training a richly parameterized logistic regression model using the data preprocessed in the (1) and (2) components of OASIS-AD. The second step consists of refining the voxel-level probability map generated in the first step to shrink WMH regions and smooth the probability map to reduce the false-positive rate. A flowchart of OASIS-AD is presented in Figure 2.2.1. In the next sections, we describe the OASIS-AD steps in greater detail.

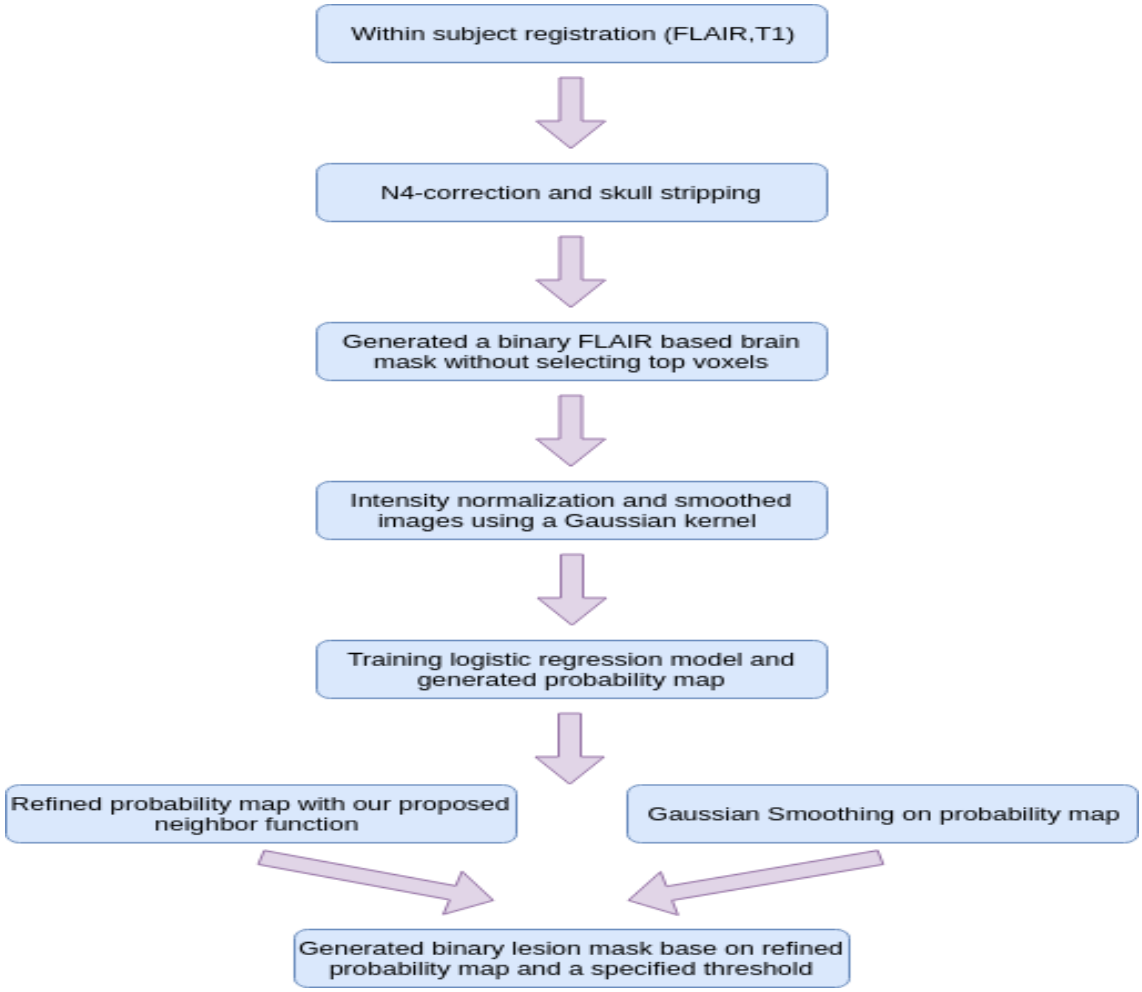

Figure 2.2.1: OASIS-AD procedure

2.2.1 Study participants

We have randomly selected a sample of 20 older individuals from our ongoing Normal Aging study previously described in (Karim et al., 2019), (Nadkarni et al., 2019). The selected sample included 20 cognitively normal study participants at the time of scanning. The average age in our sample is 81.2 (SD=7.15), with an average education equal to 14.2 years (SD=2.44), 70% of the sample are females, 85% white and 15% african american. In the next sections we describe the OASIS-AD steps in greater detail.

2.2.2 Image preprocessing

The image preprocessing used `fslr` (Muschelli et al., 2015) package in `Neuroconductor` (Muschelli et al., 2018), a comprehensive `R` environment for imaging processing tools. The `fslr` package wraps the FMRIB Software Library (FSL 5.0) (https://fsl.fmrib.ox.ac.uk/fsl) into the `R` language. The pre-processing steps were applied in the following order:

1. Within-subject coregistration of the T1-weighted image to the T2-FLAIR image.
2. Apply N4-bias-correction (Tustison et al., 2010) to the registered T1-weighted image.
3. Conduct skull stripping using FSL BET (Brain Extraction Tool) (Smith, 2002) on the registered and N4 corrected T1-weighted image.
4. Erode the brain mask with a default $5 \times 5 \times 5$ kernel box.

Eroding a binary mask, A, with a kernel, B, centered at C consists of moving B by sliding its center C over all voxels in A. If all voxels in B are contained in A then the location of the center C is labeled as 1; otherwise, it is labeled 0 (erosion) (Haralick et al., 1987). The fslerode package in fslr (Muschelli et al., 2015) was used for the erosion procedure.

2.2.3 Intensity normalization

Using a method similar to the one used by Shinohara et al. (2012), images intensities for both T1-weighted and T2-FLAIR images were normalized as follows:

$$f_i^N(v) = \frac{f_i(v) - \mu_{i,M}}{\sigma_{i,M}} \, ,$$

where $\mu_{i,M}$ and $\sigma_{i,M}$ are the mean and standard deviation of the preprocessed image intensities for subject i from modality M. Note that Shinohara et al. (2012) used the normally appearing white matter (NAWM) as the reference set for normalization, which would require at least partial segmentation of NAWM. Here we avoid this problem by using the entire brain as reference.

2.2.4 Smoothed volumes

Similarly to the original OASIS approach, we used smoothed volumes by applying Gaussian kernel smoothers both to the T1-weighted and T2-FLAIR images. Two 3D Gaussian filters with window sizes of 10 and 20 mm, respectively were used to captures local inhomogeneity patterns that were not accounted for the N4 correction. We denote by $GM_i^N(v,k)$ the smoothed volume for subject i, image modality M, and kernel size k at voxel v. We fit models that include these smoothed volumes as well as models that do not (labeled reduced models), as the aggressive smoothing might actually remove subtle differences specific to the WM/GM boundary, which could further induce classification bias.

2.2.5 Logistic regression model

The OASIS-AD model includes coefficients for intensities from the FLAIR and T1 as well as smoothed intensities from those images and interaction terms between those terms. It should be noted that OASIS-AD is flexible and able to handle more image modalities, depending on the specific application and study data. To account for the interaction among different modalities, two logistic regression models were used here: **M1**, a full model based on OASIS and all the image modalities and **M2**, a reduced model. The **M1** model for the probability that a voxel v for study participant i is in WMH is:

$$\begin{aligned}\mathbf{M1}: \operatorname{logit}\big(P\{W_i(v)=1\}\big) = \beta_0 &+ \beta_1 * \mathrm{FLAIR}_i^N(v) + \beta_2 * \mathrm{GFLAIR}_i^N(v,10) \\ &+ \beta_3 * \mathrm{GFLAIR}_i^N(v,20) + \beta_4 * \mathrm{T1}_i^N(v) + \beta_5 * \mathrm{GT1}_i^N(v,10) \\ &+ \beta_6 * \mathrm{GT1}_i^N(v,20) + \beta_7 * \mathrm{FLAIR}_i^N(v) * \mathrm{GFLAIR}_i^N(v,10) \\ &+ \beta_8 * \mathrm{FLAIR}_i^N(v) * \mathrm{GFLAIR}_i^N(v,20) \\ &+ \beta_9 * \mathrm{T1}_i^N(v) * \mathrm{GT1}_i^N(v,10) + \beta_9 * \mathrm{T1}_i^N(v) * \mathrm{GT1}_i^N(v,20).\end{aligned}$$

Model **M2** with the reduced predictors set is:

$$\mathbf{M2}: \operatorname{logit}\left(P\{W_i(v)=1\}\right) = \beta_0 + \beta_1 * \mathrm{FLAIR}_i^N(v) + \beta_2 * \mathrm{T1}_i^N(v),$$

where $\mathrm{FLAIR}_i^N(v)$ is the normalized ith voxel's FLAIR value, while $\mathrm{GFLAIR}_i^N(v,10)$ and $\mathrm{GFLAIR}_i^N(v,20)$ are smoothed normalized ith voxel FLAIR values with Gaussian kernels

of size 10mm and 20mm, respectively. Notation for the other modalities follows the same convention.

2.2.6 Probability map refinement

The logistic regression models introduced in Section 2.2.5 are used to produce an initial probability map for WMH at the voxel level. This probability map is then refined to reduce the false positive detection rate using two additional techniques: *Nearest Neighbor Refinement* and *Gaussian filter Refinement* to remove false positives. We describe these in the next two sections.

2.2.6.1 Nearest Neighbor Refinement The Nearest Neighbor Refinement (NNR) consists of first, applying the `FAST` (Zhang et al., 2001) algorithm, a popular brain tissue segmentation based on T1-weighted images. The `FAST` algorithm provides an estimated probability that the voxel v is in white matter, p^v_{wm}, gray matter, p^v_{gm}, CSF, p^v_{csf}, respectively. The sum of p^v_{wm}, p^v_{gm} and p^v_{csf} is equal to 1 for every voxel v. From these estimated tissue probability maps we estimate the tissue type of voxel v, denoted by T_v, as the tissue with the highest probability at voxel v. Using the logistic models in Section 2.2.5 we generate a probability that each voxel v is in WMH and denote it by P^v_{wmh}. We denote by N_v the 6 nearest neighbors (6NN) of voxel v. The idea is to use information from the neighboring voxels to reduce "speckling", the phenomenon where a few isolated voxels are identified as WMH when they should not be. Below we provide the detailed algorithm.

The algorithm starts with voxels whose estimated probability by `FAST` of being in white matter is 1, $p^v_{wm} = 1$, and whose 6NN are all estimated to be in white matter by `FAST`, $T_v^{6NN} = wm$. Here, the last equality indicates that all entries of the six-dimensional vector T_v^{6NN} are estimated to be white matter by `FAST`. For these voxels the estimated probability of the voxel being in WMH is exponentially reduced by simply raising the estimated probability of the voxel being in WMH using the logistic models to the power 10, $P^{rv}_{wmh} = (P^v_{wmh})^{10}$. The net effect is to substantially reduce the estimated probability of this type of voxel to be in WMH. The second option is when the voxel is estimated by `FAST` to be in white matter,

Algorithm 1 Nearest Neighbor Refinement (NNR)

Input: T_v, tissue type for voxel v estimated by `FAST`

T_v^{6NN}, tissue type set for the 6NN of voxel v estimated by `FAST`

p_{wm}^v, probability of voxel v being in white matter estimated by `FAST`

p_{wm}^{6NN}, probability set for the 6NN of voxel v estimated by `FAST`

P_{wmh}^v, probability of voxel v being WMH estimated by logistic models

Output: P_{wmh}^{rv}, probability of voxel being WMH estimated using NNR

1: **procedure** NNR(v)
2: **if** $p_{wm}^v = 1$ and $T_v^{6NN} = wm$, **then**
3: $P_{wmh}^{rv} = (P_{wmh}^v)^{10}$
4: **else if** $T_v = wm$ and $T_v^{6NN} \neq wm$, **then**
5: $P_{wmh}^{rv} = (P_{wmh}^v)^{\text{average}(p_{wm}^{6NN})}$
6: **else**
7: $P_{wmh}^{rv} = P_{wmh}^v$
8: **return** P_{wmh}^{rv}

$T_v = wm$, but not all its 6NN are estimated to be in white matter, $T_v^{6NN} \neq wm$. The last inequality indicates that at least one of the 6NN of the voxel v is not estimated to be in white matter by `FAST`. In this case, the estimated probability for the voxel to be in WMH is increased by raising it to the power average(p_{wm}^{6NN}), which is the average of the estimated probabilities for the voxel to be in white matter by `FAST`. The average of these probabilities is a number less than one, indicating that the probability will be increased. The probability is increased more when there are more neighbors that are not estimated to be in white matter and when the estimated probabilities of these neighbors are further from 1, indicating increased probability that the voxels are not actually in the white matter. Both of these choices of powers were found empirically to work well and were validated using training/test data. If neither of these conditions are satisfied than the probability map obtained from the logistic models remains unchanged, $P_{wmh}^{rv} = P_{wmh}^v$.

2.2.6.2 Gaussian Filter Refinement

Once the NNR procedure is applied we apply a 3D Gaussian filter on the generated probability maps using the following sequence of operations: (1) create an eroded brain mask, (2) fill in the voxels in the eroded brain mask with the WMH probabilities estimated in Section 2.2.6.1, and (3) apply a 3D Gaussian filter of size $5 \times 5 \times 5$ mm to the probability map on the eroded brain.

2.2.7 Binary segmentation and evaluation metrics

After creating the probability maps, a threshold value needs to be identified to classify voxels into classes. We use an approach proposed by Valcarcel et al. (2018), who proposed to use multiple threshold candidates and selected the optimal threshold based on the performance on the training set. We used the Dice Similarity Coefficient(DSC) (Dice (1945)) as the evaluation metric for selecting the optimal threshold.

Results were compared with manual segmentation performed by an experienced neuroradiologist, which provided the gold standard. The manual tracings of WMH were performed on 5 contiguous slices on the T2-FLAIR scans, the same for each subject. Models were compared in terms of the following metrics: (1) number of true positive voxels (TP), (2) number of false-positive voxels (FP), (3) number of true negative voxels (FN), and (4) number of false negative voxels (FN). We computed four additional combined metrics commonly used for prediction performance evaluation (Goutte and Gaussier, 2005): (1) accuracy, defined as ACC = (TP+TN)/ (TP+FP+FN+TN), (2) positive predictive value, defined as PPV = TP/(TP+FP); (3) true positive rate, defined as TPR = TP/(TP+FN), (4) false positive rate, defined as FPR = FP/(FP+TN), and (5) dice similarity coefficient, defined as DSC = 2TP/(2TP+FP+FN) as well as 95% confidence interval (CI) computed using bootstrap. We also included the receiver operating characteristic curve (ROC curve), the precision-recall curve (PRC), and the area under these two curves (AUC) (Davis and Goadrich, 2006).

2.2.8 Comparison with other methods

We compared OASIS-AD with four other methods: OASIS (developed for MS lesion segmentation), MIMOSA (Valcarcel et al., 2018), the lesion segmentation tool (LST) (Schmidt,

2017), and the fuzzy connected algorithm of Wu et al. (2006) and labeled as fuzzy-c. our study participants sample described in Section 2.2.1.

2.3 Results

Data including 20 subjects were randomly split into training (15 study participants) and testing (5 study participants) sets; models were trained on training data set and compared in terms of their performance on the testing data set. The proposed methods was fit using a `R` package "OASISAD", which was created for this dissertation. All analyses were conducted in `R`(Team et al., 2013).

2.3.1 OASIS-AD models comparison

We start by first evaluating the various types of the OASIS-AD model. Table 2.3.1 provides results for all model combinations considered, where the first column provides the label, while the second column provides the type of analysis conducted. For example, M2-NG is the OASIS-AD model using the logistic model M2 introduced in Section 2.2.5 combined with the NNR algorithm introduced in Section 2.2.6.1 followed by GFR algorithm introduced in Section 2.2.6.2. The acronym for this model could be OASIS-AD-M2-NG, but this is way too complex and we will use the M2-NG shortcut for presentation purposes while understanding that all these models have the OASIS concept at the core with various refinements added to the resulting probability masks. The third column in Table 2.3.1 provides the optimal threshold obtained during training, while the fourth and fifth columns display the corresponding DSC and FPR on the test data.

Results indicate that the $M1$ model series (i.e., full models) outperforms the corresponding $M2$ series models (higher DSC and better FPR), but the differences are not very large. Taking into account that the $M2$ series models do not use smooth volumes, which can be time intensive on large datasets, we consider that the $M2$ series models provide an excellent first line approach for WMH segmentation. The M1-G model achieves the highest DSC (0.78),

Table 2.3.1: OASIS-AD models information

OASIS-AD	Techniques	Optimal Threshold	DSC	FPR
M1	**M1**	0.17	0.72	0.017
M1-G	**M1** + GFR	0.20	0.79	0.011
M1-NG	**M1** + NNR + GFR	0.17	0.74	0.011
M1-GN	**M1** + GFR + NNR	0.21	0.76	0.008
M2	**M2**	0.13	0.70	0.024
M2-G	**M2** + GFR	0.14	0.77	0.017
M2-NG	**M2** + NNR + GFR	0.13	0.72	0.016
M2-GN	**M2** + GFR + NNR	0.16	0.74	0.013

though it has a slightly higher FPR than the M1-NG model (0.009 compared to 0.007).

2.3.2 Comparisons with other models

Table 2.3.2 compares results for the best OASIS-AD model (M1-G) with the four other methods: OASIS, MiMOSA, LST and fuzzy-c, and Table A.1.1 in Appendix compares results for all the OASIS-AD models with other methods. For the fuzzy-c method proposed by Wu's (Wu et al. (2006)) we only have the binary brain masks and not the probability map. Therefore, it is not possible to compute the AUCs for fuzzy-c. The OASIS-AD (M1-G) model has the highest DSC at 0.78, with a 95% CI equal with (0.77, 0.79), Both MIMOSA and LST being close in second place (DSC=0.71, 95% CI: (0.70, 0.77) and DSC=0.76, 95% CI: (0.75, 0.80) respectively). The ROC-AUC (0.97) and ROC-PRC (0.86) for the M1-G model are substantially better than for MIMOSA (0.87 and 0.77, respectively) and LST (0.87 and 0.77, respectively.)

Figure 2.3.1 displays the ROC and PRC for the four models OASIS-AD (M1-G), OASIS, MIMOSA, and LST, and Figure A.2.1 in Appendix displays the ROC and PRC for all the models except fuzzy-c. The ROC curves are indistinguishable in the area of high specificity

Table 2.3.2: Performance evaluation metrics (reduced)

	ACC	PPV	TPR	FPR	DSC	ROC	PRC
M1-G	0.97(0.01)	0.85(0.03)	0.70(0.03)	0.009(0.001)	0.78(0.03)	0.97	0.86
	(0.96,0.98)	(0.83,0.88)	(0.69,0.72)	(0.008,0.01)	(0.77,0.79)		
OASIS	0.95(0.01)	0.75(0.04)	0.58(0.04)	0.014(0.002)	0.65(0.04)	0.92	0.74
	(0.94,0.96)	(0.75,0.8)	(0.58,0.62)	(0.012,0.015)	(0.64,0.69)		
MIMOSA	0.96(0.01)	0.94(0.02)	0.58(0.04)	0.002(0.001)	0.71(0.04)	0.87	0.77
	(0.96,0.97)	(0.93,0.97)	(0.56,0.64)	(0.001,0.003)	(0.70,0.77)		
LST	0.97(0.01)	0.83(0.05)	0.72(0.04)	0.012(0.005)	0.76(0.03)	0.87	0.77
	(0.96,0.97)	(0.83,0.86)	(0.71,0.76)	(0.010,0.013)	(0.75,0.8)		
fuzzy-c	0.95(0.002)	0.88(0.13)	0.51(0.13)	0.018(0.015)	0.62(0.11)	NA	NA
	(0.94,0.96)	(0.85,0.89)	(0.50,0.52)	(0.017,0.019)	(0.61,0.63)		

Data is presented as mean (standard deviation) and 95% CI

(specificity>0.99), with the M1-G model performing slightly better. However, as specificity is allowed to be smaller (moving right on the 1-Specificity x-axis) the ROC of the OASIS-AD model is substantially better than for the other models. This indicates that small changes in specificity can lead to much larger improvements in sensitivity for the OASIS-AD model compared to the competing models. Both MIMOSA and LST seem to be tuned specifically for high specificity, whereas OASIS has higher sensitivity for specificity areas that are not of practical interest. A similar result can be noted for the PRC in the left panel of Figure 2.3.1.

2.3.3 One slice comparison among models: case study

Figure 3.6.2 showing true positives, false positives and false negatives color coded, compares the WMH segmentation results using two OASIS-AD methods (M1-G shown in panel C and M1-GN shown in panel D) with OASIS (panel E), MIMOSA (panel F), and LST (panel G), and fuzzy-c (panel H). Results are shown on one slice of a random subject from

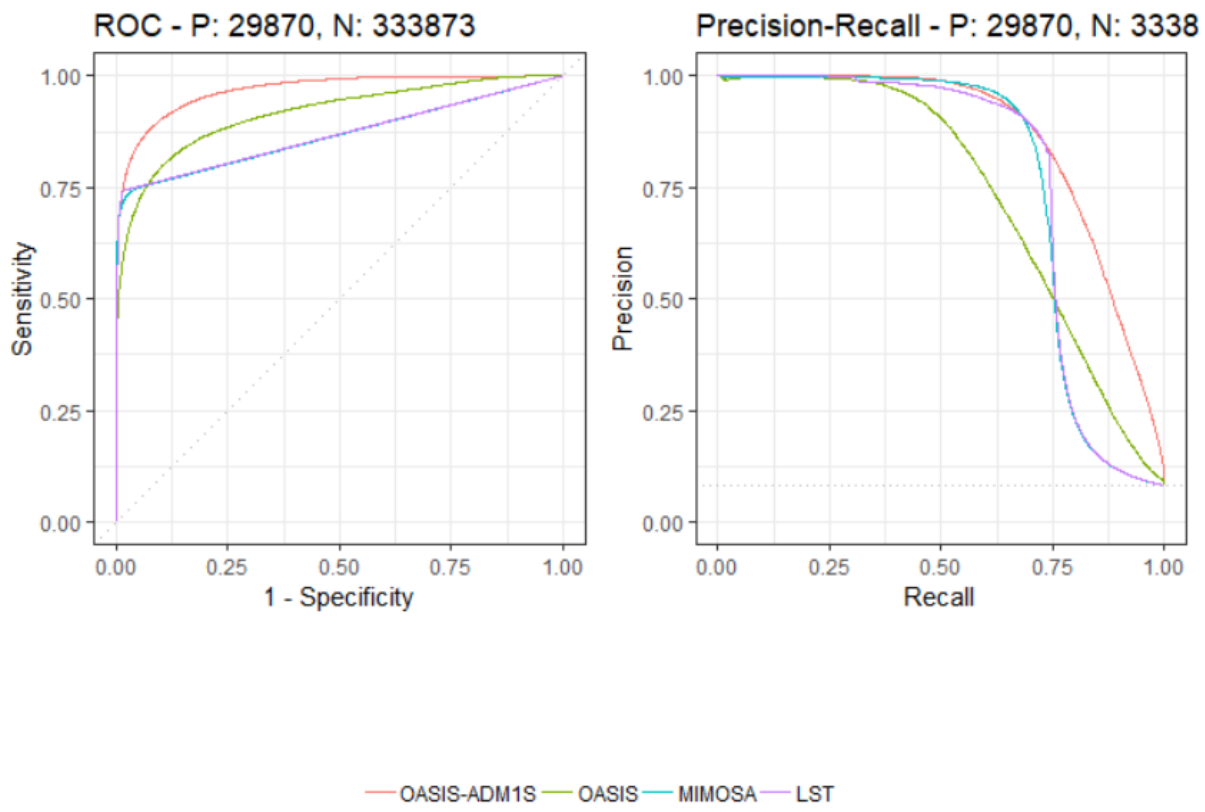

Figure 2.3.1: ROC and PRC of models(reduced)

the training data. The corresponding FLAIR slice is shown in Panel A, while the manual segmentation of WMH is shown in Panel B. This slice contains both large and small contiguous WMHs regions and results indicate the good performance of both OASIS-AD approaches. The MIMOSA mask also looks very good, with slightly more speckling. The LST and OASIS estimators seem to contain many more spatially distributed false positive voxels, which may indicate a substantially different trade-off of false positives. Indeed, while the FPR was comparable between OASIS-AD and OASIS and LST, it seems that the false positives for OASIS-AD tend to cluster close to the true positives, whereas for the other two methods they are spread in areas that do not contain WMH. The fuzzy-c mask seems to

be slightly conservative, misses important WMH clusters, and falsely identifies some WMH close to the cortical surface.

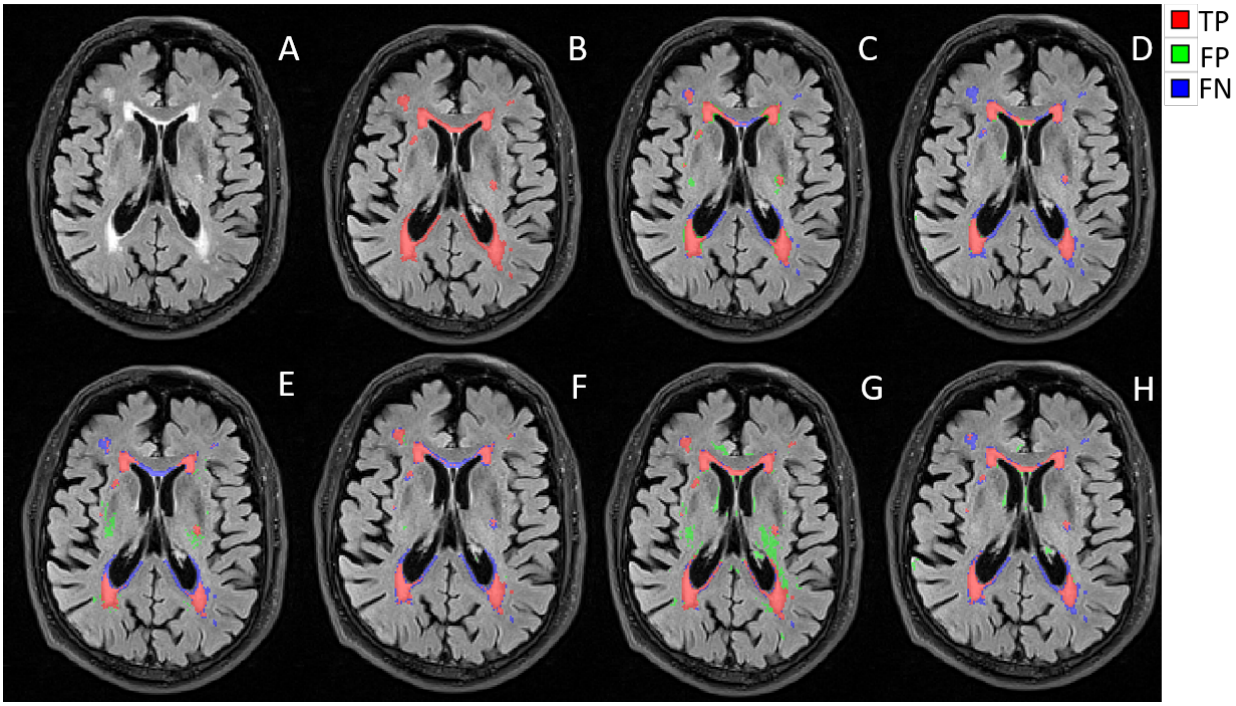

Figure 2.3.2: Case study: **A**: FLAIR slice, **B**: manual, **C**: M1-G, **D**: M1-GN, **E**: OASIS, **F**: MIMOSA, **G**: LST, **H**: fuzzy-c.

2.4 Conclusions

We introduced OASIS-AD, a class of models designed to refine OASIS (Sweeney et al., 2013), an MS lesion segmentation approach for WMH in older adults with AD. OASIS-AD performed well in comparison with existing methods. OASIS-AD provides an interpretable solution based on logistic regression combined with two map refinement techniques designed to reduce the false-positive rate. OASIS-AD is a significant improvement over OASIS both in terms of modeling techniques, which are adapted for the specific problems raised by WMH segmentation and in terms of segmentation performance. OASIS-AD has three major advantages that are worth emphasizing. First, the logistic-based approach is highly flexible

and it allows the use of any combination of multi-modal inputs, easy expansion of the predictor space, non-linearity, and potential interaction effects. Moreover, traditional methods for quantifying the relative importance of existing or new predictors can provide powerful insights into what and how new modalities and features are actually contributing to improved segmentation. Second, OASIS-AD can be trained with small, moderate, and large sample sizes, making it a very useful first-line segmentation approach that can be easily deployed in new environments or sub-disease types. Third, and probably most importantly, OASIS-AD is easy to generalize and interpret because it is based on a logistic regression model that accounts for the intensity of voxels in various disease tissues across image modalities.

3.0 Multivariate Image-on-scalar Regression via Interpretable Regularized Reduced Rank Regression

3.1 Introduction

In this project, we used the cohort in Cohen et al. (2013), which included both PET and sMRI images of patients in an Alzheimer's Disease study, and patients' demographics, psychosocial and cardiovascular measures. We developed a multivariate image-on-scalar regression that used PET as outcomes and patients' scalar measures as predictors. To overcome the high-dimension of images, we restricted our multivariate regression via a low-rank factorization of the parameter matrix. Low-rank factorization is a straightforward dimension reduction technique that can be used with high-dimensional data to introduce parsimony, resulting in theoretical and computational benefits. Reduced-rank regression (Reinsel and Velu, 1998) is a popular tool for conducting regression analyses with multivariate outcomes that utilize potentially low-rank structures of coefficient matrices to account for relationships among response variables. Several regularized reduced-rank regression procedures have been proposed that combine reduced rank-regression with regularizing penalties to facilitate parameter estimation and model selection with high-dimensional data (Chen and Huang, 2012; Chen et al., 2012, 2013; She and Chen, 2017). However, there are two major limitations in the use of existing regularized reduced-rank regression methods for the analysis of image response data. First, to the best of our knowledge, no existing method simultaneously accounts for the spatial smoothness, functional and structural grouping, and sparsity inherent with image response data. Second, regularization in existing methods is built on a subspace after dimension reduction. This leads to a lack of interpretability since either the rank factorization is not unique or the subspace lacks the same structure as the original data. For example, in our motivating application, voxels that are in close proximity or in a common structural group in the brain (i.e. image response variable space) are not necessarily the same distance apart or in the same group in the reduced-rank subspace.

To overcome these limitations, in this Chapter we propose interpretable reduced-rank re-

gression (IRRR) as a method for image-on-scalar regression. The method uses a fused sparse group lasso penalty after dimension reduction, which reduces the size of the high-dimensional model while regularizing based on spatial smoothness, structural and functional grouping, and sparsity. Many different group structures can be used based on known biological information or information that can be used to increase interpretablity of results. For example, by specifying group structures based on regions of interest (ROI), the procedure can conduct a voxel-wise analysis that fully utilizes all image information while producing results that can be interpreted as an ROI-wise analysis. The penalty includes an encoder-decoder to enable it to be formulated on the reduced-rank space, but maintain biological interpretation and regularize on the image space.

3.2 Method

3.2.1 Model

Let $\Re^{n\times m}$ be the space of $n\times m$ real-valued matrices. We observe images with m response variables (i.e. voxels) from n independent subjects that have been vectorized to obtain the matrix image data $Y \in \Re^{n\times m}$. Further, we assume that we observe p scalar predictors for each subject, and let $X \in \Re^{n\times p}$ represent the matrix of scalar predictors. We assume the image-on-scalar regression model

$$Y = XA + E, \tag{3.1}$$

where $A \in \Re^{p\times m}$ is a matrix of coefficients whose ijth element represents the ith scalar predictor's effect on the jth image response variable, and the elements of $E \in \Re^{n\times m}$ are independent mean-zero Gaussian random variables with variance σ^2. It is assumed that data have been centered so that no intercept appears in the model.

Ordinary least squares estimation of this model is undesirable for two reasons. First, it provides estimators that ignore the existence of any relationship among response variables, resulting in an estimator that is equivalent to conducting univariate regressions on each outcome variable individually. In practice, especially for image data, outcome variables are

highly related. Second, with high-dimensional data, some sort of dimension reduction is essential for obtaining stable and tractable estimates. To overcome these issues, we assume that the model is of reduced rank such that $r = \text{rank}(A) < \min(p, m)$. Let $A = BV^T$ be a rank factorization of A where $B \in \Re^{p \times r}$ and $V \in \Re^{m \times r}$ represent left and right singular subspace, respectively, such that

$$Y = XBV^T + E. \tag{3.2}$$

The row space of B represents the structure of X, or the scalar predictors, and the row space of V represents the structure of Y, or the image response variables. Our goal is to understand associations between the predictors and image responses by estimating A via B and V. It should be noted that the rank factorization is not unique since, for any orthogonal $Q \in \Re^{r \times r}$, $A = BV^T = BQQ^TV^T$. The proposed estimator circumvents this obstacle through a two-step procedure that provides a consistent estimator of A without additional constraints.

3.2.2 IRRR: Interpretable regularized reduced-rank regression

Similar to some existing regularized reduced-rank regression procedures, we will take a two-step approach to estimation that first estimates B, then V. This approach has two favorable characteristics. First, estimating V conditional on an estimate of B mitigates potential identifiability issues without needing to introduce geometric constraints. Second, separating the estimation of B and V provides a divide-and-conquer type of approach that reduces the size and complexity of any individual optimization. The innovative question considered in this article is in analyses with image response variables while accounting for the complex structure inherent with image data. The two-stage procedure enables us to isolate this complexity to the estimation of V. It also allows us to utilize existing methods for reduced-rank regression with potentially high-dimensional predictors to estimate B. For example, a consistent estimate of B can be obtained using methods such as that considered by Ma and Sun (2014). In this subsection, we discuss the proposed novel estimator of V given an estimate $\hat{B}$ of B. The full proposed estimation procedure, including the estimation of B and inherent selection of r, is presented in Section 3.3.

There are three aspects of image data that we want to exploit in obtaining regularized estimators. Penalties will be formulated using the L_1 and L_2 matrix norms, defined for a matrix M as $|M|_1 = \sum_j \sum_k |m_{jk}|$ and $||M||_2 = (\sum_j \sum_k |m_{jk}|^2)^{1/2}$, respectfully. First, we desired a fused penalty that regularizes based on the smoothness of adjacent image response variables. Let $D \in \Re^{N_F \times m}$ be the generalized lasso representation of the fused lasso such that $\left|AD^T\right|_1$ is the sum of differences of adjacent image response variables (Tibshirani, 2011), illustrations of which are provided in the Appendix. Spatial smoothness will be accounted for by penalizing the roughness $\left|AD^T\right|_1 = \sum_{i=1}^{p} \sum_{j=1}^{m-1} \sum_{k \in \mathbb{N}_j} |A_{ij} - A_{ik}|$ where $\mathbb{N}_j$ is the set of image response variables one unit larger than the jth in any dimension. In addition to spatial smoothness, often with imaging data, either based on prior findings or on a desire to obtain more interpretable results, image response variables can be placed into group or clusters based on functional or structural networks. An interpretable solution would allow one to regularize by selecting entire groups of voxel effects. Given a set $\mathbb{G}$ of non-overlapping subsets of the m image response variables, for a $g \in \mathbb{G}$ with m_g elements, we define $G_g \in \Re^{m_g \times m}$ as the matrix such that $AG_g^T \in \Re^{n \times m_g}$ is the submatrix of A with columns corresponding to the elements of g. Group structure will be accounted for by incorporating the group lasso penalty $\sum_{g \in \mathbb{G}} m_g^{1/2} \left|\left|AG_g^T\right|\right|_2 = \sum_{i=1}^{p} \sum_{g \in \mathbb{G}} m_g^{1/2} (\sum_{j \in g} A_{ij}^2)^{1/2}$ (Yuan and Lin, 2006) . Lastly, we will allow for sparsity among voxels within groups, as well as sparsity among voxels not included in a group, through the lasso penalty $|A|_1 = \sum_{i=1}^{p} \sum_{j=1}^{m} |A_{ij}|$ (Tibshirani, 1996).

Given an estimate $\hat{B}$ of B, we formulate an estimator of V that uses the linear operator $\hat{B}$ as an encoder decoder. Let $X_{\hat{B}} = X\hat{B}$, so that Equation (3.2) be written at $Y = X_{\hat{B}}V + E$. The regression coefficient subspace of $\Re^{n \times m}$ of rank r matrices with left singular subspace $\hat{B}$ can be represented as $\hat{B}V^T$, $V \in \Re^{m \times p}$. Rather than formulating penalties on A, we formulate them on this subspace and replace A with its projection $\hat{B}V^T$. Formally, given tuning parameters $\lambda_1, \lambda_2, \lambda_3 > 0$, which control the degree of regularization through the lasso, fused lasso and group lasso penalties, respectively, the IRRR estimator is defined as $\hat{A} = \hat{B}\hat{V}^T$ where

$$\hat{V} = \underset{V \in \Re^{m \times r}}{\arg\min} \frac{1}{2} \|X_{\hat{B}}V - Y\|_2^2 + \lambda_1 |\hat{B}V^T|_1 + \lambda_2 |\hat{B}V^T D^T|_1 + \lambda_3 \sum_{g \in \mathbb{G}} m_g^{1/2} ||\hat{B}V^T G_g^T||_2. \quad (3.3)$$

This formulation allows us to estimate within the tractable reduced-rank subspace while regularizing based on penalties that are interpretable on the high-dimensional image response space.

3.3 Estimation Procedure

3.3.1 Two step estimation algorithm

We propose a two-step estimation procedure, which is formally defined in Algorithm 2. The first step involves the estimation of B and is a modification of the procedure considered by Ma and Sun (2014). As opposed to Ma and Sun (2014), who considers high dimensional p under sparsity, we are concerned with moderate p of potentially highly correlated predictors that are selected for their biological relevance; subsequently all are expected to be associated with some image response variables. We replace their lasso penalty with a ridge penalty (Hoerl and Kennard, 1970), which can be efficiently solved using the algorithm of Friedman et al. (2010). It should be noted that Algorithm 2 can be easily adjusted to include the case of high-dimensional sparse predictors; a discussion of this extension is provided in Section 3.7. The key step in the algorithm is solving Equation (3.3) an outline of an algorithm for which is given in the following subsection, with technical details provided in the Appendix. In Algorithm 2, the function $\rho_B(B;\lambda) = \lambda||B||_2^2$ is the ridge penalty, the function $\rho_V(V;\lambda)$ is the fused sparse group penalty found in Equation (3.3) and, to simplify presentation, we adopt a slight abuse of notation and let λ represent general tuning parameters.

The algorithm depends on several parameters. The ridge and sparse fused group lasso regressions depend on tuning parameters, which can be selected through 5-fold cross-validation. We estimate the rank $r = \text{rank}(A)$ using the method of (Bunea et al., 2011) and the standard deviation of the errors as $\hat{\sigma} = \text{median}\left\{\sigma(Y)\right\}/\sqrt{\max(n,m)}$ (Ma and Sun, 2014).

Algorithm 2 Two step estimation algorithm of A

Input:Vectorized image data Y, scalar data X, estimated rank r, noise level σ, ridge regularization on $\hat{B}$: $\rho_B(\cdot;\lambda)$, fused sparse group lasso on $\hat{V}$: $\rho_V(\cdot;\lambda)$

Output: $\hat{A}$

1: Compute $P = X(X^TX)^-X^T$, where $(X^TX)^-$ is Moore-Penrose pseudo-inverse

2: Compute right singular subspace of PY by singular value decomposition with first r singular vectors, denoted as V_0

3: Ridge regression:

$$B_1 = \underset{B\in\Re^{p\times r}}{\arg\min}\, ||YV_0 - XB||_2^2 + \rho_B(B;\lambda)$$

4: Compute the left singular vectors of XB_1, denoted as U_1

5: Compute the right singular vectors of $U_1U_1^TY$, denoted as V_1.

6: Ridge regression:

$$B_2 = \underset{B\in\Re^{p\times r}}{\arg\min}\, ||YV_1 - XB||_2^2 + \rho_B(B;\lambda)$$

7: Fused sparse group penalized regression:

$$V_2 = \underset{V\in\Re^{m\times r}}{\arg\min}\, ||Y - XB_2V^T||_2^2 + \rho_V(V;\lambda)$$

8: Compute estimation of A: $\hat{A} = B_2V_2^T$

3.3.2 Alternating direction method of multipliers (ADMM) solution

To solve Equation (3.3), we first represent it in a more computationally amenable form. Then, we use alternating direction method of multipliers (ADMM) (Boyd et al., 2011) to obtain a numeric solution. The ADMM has the ability to handle complicated penalty structures, such as the one encountered in Equation (3.3), that cannot be separated into a sum of functions of the elements of V. Such penalties structures are not amenable to many other common approaches, such as coordinate descent and accelerated gradient.

We begin by noting that Equation (3.2) can be expressed as $Y^v = X_B^vV^v + E^v$ where $Y^v = \text{vec}(Y)$, $X_B^v = [I_m \otimes (X\hat{B})]$, $V^v = \text{vec}(V)$, and $E^v = \text{vec}(E)$. Next, we represent the penalty term as the sum of L_2-norms. This can be done by recognizing that the trivial relationship

$|v| = \sqrt{v^2}$ enables the penalty function to be written as the sum of $N = N_L + N_F + N_G$ L_2-norms (Beer et al., 2019): $N_L = pm$ from the lasso penalty, N_F from the fused penalty where, for a 3-dimensional image of dimension $m_1 \times m_2 \times m_3$ such that $m = m_1 m_2 m_3$, $N_F = p(3m - m_1 m_2 - m_1 m_3 - m_2 m_3)$, and $N_G = p|\mathbb{G}|$ from the group penalty where $|\mathbb{G}|$ is the number of groups. Thus, Equation (3.3) can be represented

$$\hat{V}^v = \underset{V^v \in \Re^{(mr)}}{\arg\min} \frac{1}{n} \|X_B^v V^v - Y^v\|_2^2 + \sum_{\ell=1}^{N} \lambda_\ell ||K_\ell V^v||_2, \tag{3.4}$$

where for $\ell = 1, \ldots, N_L$, K_ℓ is the ℓth row of $I_m \otimes \hat{B}$ and $\lambda_\ell = \lambda_1$, for $\ell = N_L+1, \ldots, N_L+N_F$, K_ℓ is the $(\ell - N_L)$th row of $D \otimes \hat{B}$ and $\lambda_\ell = \lambda_2$, and for $\ell = N_L + N_F + 1, \ldots, N$, $K_{\ell+N_L+N_F}$ is the matrix $G_{g_\ell} \otimes \hat{B}$ and $\lambda_\ell = \lambda_3 m_{g_\ell}^{1/2}$ where g_ℓ is some ordering of the N_G groups in $\mathbb{G}$. Lastly, letting $K = (K_1^T \mid \cdots \mid K_N^T)^T \in \Re^{N \times mr}$ be the concatenation of the matrices K_ℓ and introducing auxiliary variables μ_ℓ, $\theta_\ell = K_\ell V^v$, and μ, θ, after initialization, the algorithm iterative updates are:

$$\begin{aligned}
V^{v(t+1)} &= \left(X^{vT} X^v + \rho K^T K\right)^{-1} \left[X^{vT} + K^T(\mu^{(t)} + \rho\theta^{(t)})\right], \\
\theta_\ell^{(t+1)} &= \left[1 - \lambda_\ell / (\rho ||\eta_\ell^{(t)}||_2)\right]_+ \eta_\ell^{(t)}, \\
\mu_\ell^{(t+1)} &= \mu_\ell^{(t)} + \rho \left(\theta_\ell^{(t+1)} - K_\ell V^{v(t+1)}\right),
\end{aligned}$$

where ρ is pre-specified step size parameter (Boyd et al., 2011), $[\cdot]_+ = \max(0, \cdot)$ and $\eta_\ell^{(t)} = K_\ell V^{v(t)} - \mu_\ell^{(t)}/\rho$. The stopping criteria of this numeric algorithm is provided in Appendix.

3.4 Theoretical Properties

In this section, we establish the consistency of $\hat{V}$ if the true B was known, then establish the consistency of $\hat{A}$ from the proposed two-step algorithm. We consider the setting where both the number of imaging variables m and the number of subjects n grow, but where the number of predictors p is fixed. The results depend on several assumptions. First, it depends on the regularity size of the design matrix. We assume $X^T X/n$ converges to a non-singular $p \times p$ matrix with maximal diagonal element d_X. Second, it depends on the sparsity and

smoothness of the parameter matrix A. We assume a fixed parameter S, which is formally defined in the Appendix. The parameter is a standard inverse measure of sparsity that is a positive function of the proportion of non-zero imaging response variables, the proportion of unequal adjacent response variables, and the proportion of non-zero groups of response variables.

Theorem 1. *Let $A^* = B^* V^{*T}$ be a rank factorization of the true coefficient matrix A^* and $\hat{V}$ be the minimizer of (3.3) given B^*. If $\max(\lambda_i) = 2Cd_X\sigma\sqrt{\log(pm)}$ for some $C > \sqrt{2}$, with probability $1 - (pm)^{1-C^2/2}$ and as $m, n \to \infty$, then*

$$||\hat{V} - V^*||_2^2 = O_p\left[\frac{\log(m)}{n}\right],$$
$$\frac{1}{n}||XB^*(\hat{V} - V^*)||_2^2 = O_p\left[\frac{\log(m)}{n}\right].$$

The results of Theorem 1, which could be of interest in their own right, can be used to establish the consistency of $\hat{A}$ from the two-stage procedure defined in Algorithm 2. We assume that an appropriate $\sqrt{n}$-consistent estimator of B has been used in the first step of the algorithm. It should be noted that, in our setting, this includes both ridge regression and least squares.

Theorem 2. *Let $A^* = B^* V^{*T}$ be a rank factorization of the true coefficient matrix and $\hat{A}$ be the estimator obtained from the two-stage procedure introduced in Section 3.3. If $\lambda_i \sim \sqrt{\log(m)}$ as $m, n \to \infty$, then*

$$||\hat{A} - A^*||_2^2 = O_p\left[\frac{\log(m)}{n}\right],$$
$$\frac{1}{n}||X(\hat{A} - A^*)||_2^2 = O_p\left[\frac{\log(m)}{n}\right].$$

The consistency in Theorem 2 was established for the fixed p, large m and n setting. A discussion about adjustments for the large but sparse p setting is provided in Section 3.7.

3.5 Simulation Study

In this section, we report results from simulation studies to evaluate the empirical properties of the proposed IRRR procedure and to compare it to existing reduced-rank regression methods that do not account for structure inherent with image data. The setting is chosen to reflect the scenario where there is sparsity and where this sparsity can be dependent on combinations of predictors and image responses. To generate a coefficient matrix $A \in \Re^{p \times m}$ for a given level of sparsity $s \in (0, 1)$, which reflects the percent of image responses with non-zero associations with any of the predictors, we begin by simulating values for its first $m \times s$ columns from a standard normal distribution and setting the remaining columns to zero. Next, the top left $[p/2] \times [(m \times s)/2]$ sub-matrix of A, which we refer to as A_0, is set to zero. Figure 3.5.1 displays a realization of A to illustrate this structure. The predictor matrix $X \in \Re^{n \times p}$ is simulated from multivariate distribution $\mathcal{N}(0, \Sigma_x)$ and Σ_x has diagonal elements 1 and off-diagonal elements ρ_X. The elements of $E \in \Re^{n \times m}$ are generated from independent standard Gaussian random variables, and outcomes are generated as $Y = XA + E$. Data are generated for $m = 200$ image response variables, $n = 100$ subjects, and varying levels of p, s and ρ_X.

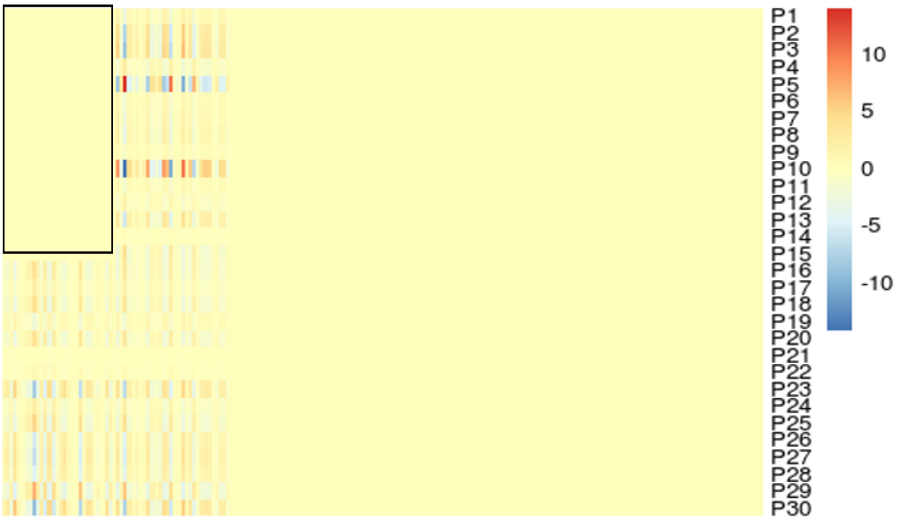

Figure 3.5.1: Illustration of a simulated coefficient matrix A

In addition to the proposed IRRR procedure, each simulated data set was also fit using

ordinary least squares (OLS) reduced-rank regression (Reinsel and Velu, 1998) and using 4 existing regularized reduced-rank regression procedures: (1) R4 - robust reduced rank regression for joint modeling and outlier detection proposed in She and Chen (2017), (2) RSSVD - the iterative procedure with sparse singular value decomposition on the regression coefficient matrix and estimated singular subspace and singular values of Chen et al. (2012), (3) RRR - the method with adaptive nuclear norm penalization of Chen et al. (2013), and (4) SRRR - the method with row-wise penalization after dimension reduction proposed in Chen and Huang (2012). These 5 existing approaches were fit using the R package "rrpack". The proposed method was fit using the provided R package "irrr", which utilizes a parallel framework and "Rcpp" (Eddelbuettel and François, 2011) to enable efficient and practical computation. Estimator performance, which is provided in Table 3.5.1, was evaluated through the average and standard deviation of mean squared error over 100 random samples per setting of 3 quantities: (1) estimation error through MSE($\hat{A}$), (2) mean prediction error through MSE(X$\hat{A}$), and (3) error in estimation of the structural zero submatrix through MSE($\hat{A}_0$). As it is known that OLS reduced-rank regression performs poorly when the dimension is larger than the sample size, it is not surprising that OLS displayed higher error compared to the 5 regularized procedures in nearly all settings. The one exception is RSSVD for highly correlated predictors, which is unable to accurately induce sparsity when there is little distinction among predictors. The proposed IRRR procedure has lowest MSE for all settings. This can be attributed to the fact that all other regularized procedures regularize on the reduced rank space, while IRRR regularizes on the image response space to maintain interpretability. The simulation settings reflect the plausible scenario where the sparsity structure of the image response and reduced rank spaces are different. Existing methods, which regularize on row-wise and column-wise reduced-rank subspaces separately, will inaccurately regularize arbitrary sparse structure within some non-sparse structure.

3.6 Analysis of PET Data

For decades, researchers have observed links between cardiovascular and Alzheimer's diseases. However, the exact nature of these connections, in particular how vascular measures are associated with anatomical symptoms of Alzheimer's disease, are unknown (Tublin et al., 2019). Clinicians and researchers utilize PET to quantify the severity of Alzheimer's disease by measuring the accumulation of the β-amyloid peptide (Aβ) in different location of the brain. The goal our the analysis considered in this section is to better understand the association between Aβ accumulation in the brain, as quantified in PET scans, with vascular measures that are easily an commonly recorded by clinicians, which are predictive of preclinical cardiovascular disease. We consider data from a study of older adults (Cohen et al., 2013) that consist of PET images in $n = 55$ older adults along with 5 vascular measures: resting pulse rate, diastolic blood pressure, systolic blood pressure, body mass index(BMI) and wait hip circumference ratio(WHR). The mean participant age was 79.32 years with a standard deviation of 6.41 years.

In this study, our analysis considers data from 5 regions of interest (ROIs): 11856 voxels comprising the anterior cingulate (ACG), 17401 voxels comprising the insula (INS), 28288 voxels comprising the orbito frontal (OBF), 13539 comprising the posterior cingulate (POC), and 21743 voxels comprising precuneus (PRE) (Cohen et al., 2013). The group penalty was formulated from these $G = 5$ groups. The top row of Figure 3.6.2 displays the location of the regions of interest in the brain. Since measurements among scalar predictors are quite different, we center each predictor and scale them by their standard deviation. For image response variables, we center them without scaling as they have already been preprocessd.

Results of exploratory univariate analyses can be found in the Appendix. The estimated coefficient matrix $\hat{A}$ from the proposed IRRR procedure is displayed in Figure 3.6.1, while the second to the fifth rows of Figure 3.6.2 display the estimated coefficients mapped onto locations of the brain. We found that associations are not present in INS, there are weak signals in OBF and relatively stronger signals in ACG, POC and PRE. Increased image intensity is a measure of increased Aβ accumulation within the brain, which is a physiological underpinning Alzheimer's behavioral manifestations. Past studies have found associations

between BMI and Aβ accumulation (Hsu et al., 2016), while our study found no association. It should be noted that we conducted a multivariate analysis so that the effect of BMI on Aβ accumulation is conditional on other variables. In particular, it is conditional on WHR, which is positively correlated with BMI. Our findings, where WHR is positively associated with Aβ accumulation in the ACG, POC and PRE conditional on BMI provides biological evidence that supports previous findings in which increased abdominal fat was found to be associated with increased risk for dementia (West and Haan, 2009). Hypertension is associated with Alzheimer's disease and one would initially expect a positive association with Aβ accumulation (Tublin et al., 2019). However, it is interesting to note that, in our analysis, systolic and diastolic blood pressure were found to be negatively associated with Aβ accumulation in some voxels, and not associated with others. This is most probably due to a selection bias. In order to be in our study, participants were required to be dementia free and healthy enough to participate in the imaging study. This inherently excludes individuals with both high blood pressure and Aβ accumulation, who would be either be demented, deceased or too ill to participate in our study. The proposed estimation procedure's ability to incorporate regularity within the brain space, which is not restricted to row- and column-sparsity on the reduced space, makes it uniquely able to identify that the conditional relationship between blood pressure and Aβ accumulation is not uniform within regions of interest.

3.7 Conclusions

The proposed IRRR represents a novel approach to image-on-scalar regression after dimension reduction with possibly hundreds of thousands of response variables that regularize based on interpretable characteristics of image data. The estimator is formulated for the setting that is common in practice, including our motivating application, where p is fixed to reflect the use of a set of predictors selected for scientific interest and where m can grow at an exponential rate compared to n to reflect the large number of image response variables relative to the number of subjects. Theoretically, under the large p setting, the consistency

results established in Theorem 2 would need to be adjusted and would be rate limited by the growth of p relative to n. Numerically, IRRR utilizes ADMM to solve estimation in a complicated regularization setting. Further discussions and future researches have been described in Chapter 5.

Table 3.5.1: Simulation Results - MSE of $\hat{A}$, $X\hat{A}$, and $\hat{A}_0$ (multiplied by 100)

Parameters			IRRR	R4	RSSVD	RRR	SRRR	OLS
p	s	ρ_X	MSE($\hat{A}$)					
10	0.1	0.1	0.10(0.01)	0.27(0.09)	0.65(0.44)	0.21(0.03)	0.27(0.09)	1.17(0.07)
10	0.1	0.9	0.39(0.12)	2.23(4.92)	10.18(8.67)	1.17(0.22)	2.18(4.92)	10.01(0.64)
10	0.9	0.1	0.21(0.03)	0.29(0.12)	0.40(0.13)	0.22(0.03)	0.28(0.11)	1.18(0.08)
10	0.9	0.9	1.06(0.20)	1.79(1.25)	12.89(8.02)	1.20(0.19)	1.73(1.20)	10.10(0.73)
30	0.1	0.1	0.05(0.01)	0.12(0.07)	0.40(0.26)	0.08(0.01)	0.11(0.06)	1.56(0.07)
30	0.1	0.9	0.24(0.06)	0.83(0.54)	19.09(14.30)	0.52(0.10)	0.78(0.47)	13.91(0.62)
30	0.9	0.1	0.08(0.01)	0.12(0.06)	0.20(0.07)	0.08(0.01)	0.12(0.06)	1.57(0.07)
30	0.9	0.9	0.49(0.09)	0.81(0.50)	22.02(13.81)	0.54(0.09)	0.76(0.43)	14.00(0.63)
p	s	ρ_X	MSE($X\hat{A}$)					
10	0.1	0.1	0.97(0.14)	2.63(0.92)	6.23(4.21)	2.07(0.31)	2.58(0.94)	11.03(0.79)
10	0.1	0.9	1.11(0.23)	3.21(4.76)	13.53(10.43)	2.13(0.31)	3.15(4.78)	11.03(0.79)
10	0.9	0.1	2.06(0.31)	2.82(1.11)	3.86(1.24)	2.15(0.33)	2.76(1.08)	11.15(0.86)
10	0.9	0.9	1.97(0.31)	2.75(1.26)	17.33(9.60)	2.12(0.31)	2.67(1.23)	11.15(0.86)
30	0.1	0.1	1.29(0.18)	3.39(2.00)	11.63(7.77)	2.43(0.35)	3.30(1.90)	43.07(2.41)
30	0.1	0.9	1.36(0.31)	3.41(1.72)	64.92(44.02)	2.44(0.35)	3.25(1.51)	43.07(2.41)
30	0.9	0.1	2.33(0.30)	3.38(1.55)	5.79(2.07)	2.41(0.31)	3.32(1.53)	43.25(2.24)
30	0.9	0.9	2.33(0.34)	3.34(1.55)	74.13(47.92)	2.50(0.35)	3.18(1.33)	43.25(2.24)
p	s	ρ_X	MSE($\hat{A}_0$)					
10	0.1	0.1	0.15(0.07)	0.36(0.18)	4.77(6.27)	0.30(0.12)	0.35(0.17)	1.16(0.23)
10	0.1	0.9	1.29(0.71)	13.92(99.47)	58.45(91.06)	1.91(0.92)	13.80(99.49)	9.90(1.80)
10	0.9	0.1	0.18(0.04)	0.29(0.13)	0.44(0.21)	0.22(0.05)	0.28(0.12)	1.17(0.12)
10	0.9	0.9	0.81(0.36)	1.82(1.54)	8.21(8.26)	1.21(0.63)	1.72(1.49)	10.16(1.06)
30	0.1	0.1	0.17(0.06)	0.25(0.09)	2.13(2.05)	0.22(0.07)	0.24(0.09)	1.55(0.21)
30	0.1	0.9	1.55(0.62)	2.02(0.76)	105.28(88.60)	1.74(0.62)	1.86(0.71)	13.75(1.86)
30	0.9	0.1	0.07(0.01)	0.12(0.06)	0.24(0.10)	0.09(0.01)	0.12(0.06)	1.56(0.10)
30	0.9	0.9	0.40(0.16)	0.80(0.54)	14.58(10.35)	0.53(0.25)	0.72(0.43)	14.08(0.91)

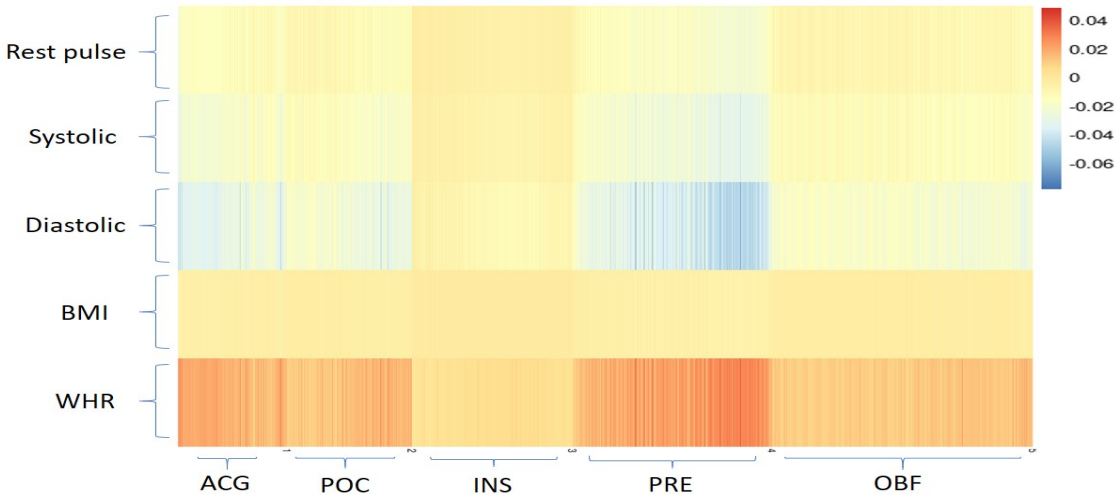

Figure 3.6.1: Estimated coefficient matrix $\hat{A}$ from the PET study.

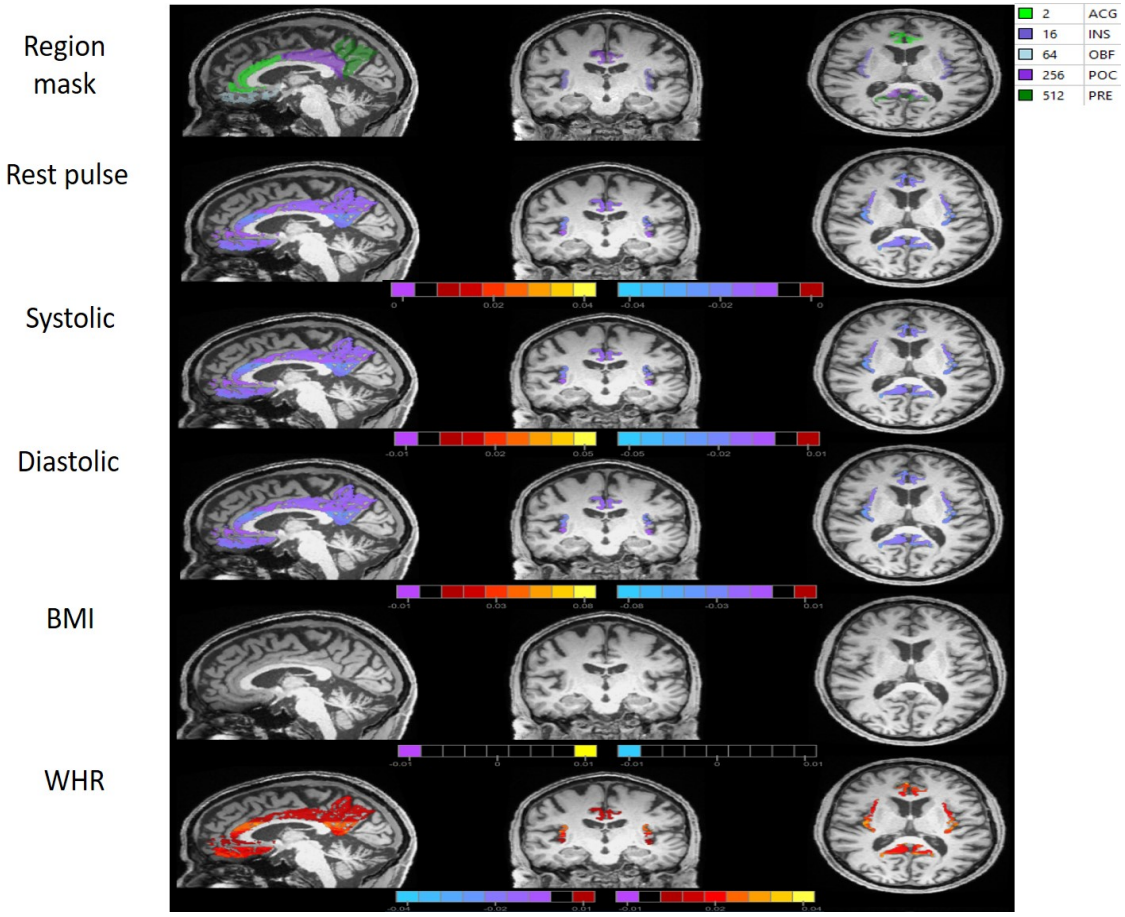

Figure 3.6.2: Location of regions of interest within the brain (1st row) and IRRR estimated regression coefficients mapped onto the brain (2nd - 6th rows) from axial (1st column), sagittal (2nd column) and coronal (3rd column) views.

4.0 ACU-Net: An Efficient Convolutional Network for Biomedical Image Segmentation

4.1 Introduction

In 1998, "Lenet-5", one of the earliest convolutional neural networks, was proposed in LeCun et al. (1998) and achieved great success for the classification of handwritten numbers on MNIST set. However, due to hardware limitations, deep neural networks (DNN) did not attract widespread attention until 2012 when AlexNet (Krizhevsky et al., 2012) won the ImageNet competition by an 11% margin. After that, both academic and IT industrial researchers developed multiple well-validated deep networks of computer vision such as Visual Geometry Group (VGG) in Simonyan and Zisserman (2014), GoogLeNet in Szegedy et al. (2015), and ResNet in He et al. (2016). Building upon these established deep network architectures, researchers developed more and more networks for specific areas or tasks. U-Net was developed in Ronneberger et al. (2015) and provides a practical deep network on training data with relative small sample size (i.e. 30 medical images and 512×512 pixels).

Recently, researchers have focused on building compact deep neural networks not limited to CNN. Wen et al. (2016) proposed a Structured Sparsity Learning (SSL) method to regularize the structures of DNNs by introducing sparse group lasso regularization, both filter-wise and shape-wise. Yu et al. (2017) assumed weight filters to be both low-rank and sparse, and split the weight matrix into the sum of a low-rank matrix and a sparse matrix, then applied several of thefamous networks listed above. Lee et al. (2019) proposed DeepTwist, a technique to compress CNNs by low-rank approximation to injected noise into weights. Kossaifi et al. (2019) proposed T-Net, a parametrizing fully convolutional network with a single high-order tensor that is different from previous layer-by-layer tensorization. Compared with the popular dropout technique in Srivastava et al. (2014), which shrunk DNNs by randomly dropping units (along with their connections) from the neural network during training, low-rank approximation and sparsity regularization provide a more interpretable approach for dimension reduction and feature selection.

In state-of-the-art biomedical image segmentation deep neural models, U-Net (Ronneberger et al., 2015) is the most famous and well-validated structure. Recent deep neural networks for biomedical image segmentation frequently use U-Net as basic structure or for comparison. Our work is also inspired by U-Net. In addition, among most recent researches on compression of deep neural networks, depth-wise separable convolutions (Howard et al., 2017), inverted residual block (Sandler et al., 2018) and squeeze-and-excitation networks (Hu et al., 2018) are proved to be very useful and popular. Thus, we proposed ACU-NET, an asymmetric compact U-Net by applying the depth-wise separable convolutions in an inverted residual block with squeeze-and-excitation to convolutional layers.

This Chapter describes the ACU-Net model in order to deliver the next generation of high accuracy efficient networks to improve biomedical imaging segmentation tasks by reducing computation cost while maintaining predictive performance. This could enable the segmentation tasks to even be performed on mobile devices in the future.

The goal of this Chapter is to optimize the trade off between accuracy and model size. To realize this we have introduced: (1) an efficient convolutional layer block design and (2) a new network architecture. We presented experiments on the normal aging cohort used in Chapter 2 of this dissertation to demonstrate the breakthrough efficacy of ACU-Net.

4.2 Method

First, we introduce U-Net architecture, which is illustrated in Fig 4.2.1. It is mainly established with:

- **Convolutional layers with ReLU (Rectified Linear Unit)** i.e. $f(x) = X^+ = max(0, x)$. Convolutional operation is sliding a convolutional filter over an input feature map. The output feature map is built by the dot products between the filters and input feature map.
- **Max-pooling layers** are operating independently on every depth slice of the input feature map and resizes it spatially, using the max function. These layers are often used to decrease the size of the input feature map.

- **Up-convolutional layers** are doubling the input feature map size.
- **Sigmoid activation function** i.e. $f(x) = 1/(1+e^{-x})$ is used in the last fully connected layer to create an output probability map.

Second, U-Net has symmetric architecture, in each convolutional layers block, the last convolutional layer is cropped and copied to corresponding up-convolutional layer block in decoder. U-Net has following properties: (1) In each convolutional layer block, it includes two convolutional layers followed by a max-pooling layer to halve the input feature maps dimension and then, double the number of channels. For example, the last feature map of first convolutional layer block is $568 \times 568 \times 64$ which is corresponding to $H(height) \times W(width) \times C(channel)$, then after a max-pooling layer, the feature map becomes $284 \times 284 \times 64$. Next, a convolutional step makes this feature map become $284 \times 284 \times 128$. (2) In the decoder part of U-Net, which is the right part of the U-Net architecture, it is symmetric to its corresponding encoder part. Thus it costs similar even higher computation cost compared with its corresponding encoder part as it concatenates the encoder part at the beginning of each decoder block.

4.2.1 ACU-Net convolutional layer block

Although U-Net is effective in biomedical imaging segmentation, it is "overweight" compared with modern compact models. To compress U-Net while maintaining its capacity, we have developed ACU-Net. Before we demonstrate ACU-Net architecture, we first introduce several techniques we have used to build ACU-Net convolutional layers block.

Depthwise separable convolution Depthwise separable convolution was proposed in Howard et al. (2017) and described in Fig 4.2.2. The classic convolutional filters, for example, with filter size $D_K \times D_K$, input channel number as M, output channel number as N in Fig 4.2.2.(a) has been decomposed into two parts: depthwise convolutional filters in (b) and pointwise convolutional filters in (c). It is called "depthwise" because this technique first looks at each channel as shown in (b), which is similar as decomposing a length M channel tensor into M length 1 tensor. Thus, this step generates a temporary output feature map with dimension $D_G \times D_G \times M$ where D_G is the spatial width and height of a square output feature map (for simplicity of illustration, we use square feature map here). Next, in order

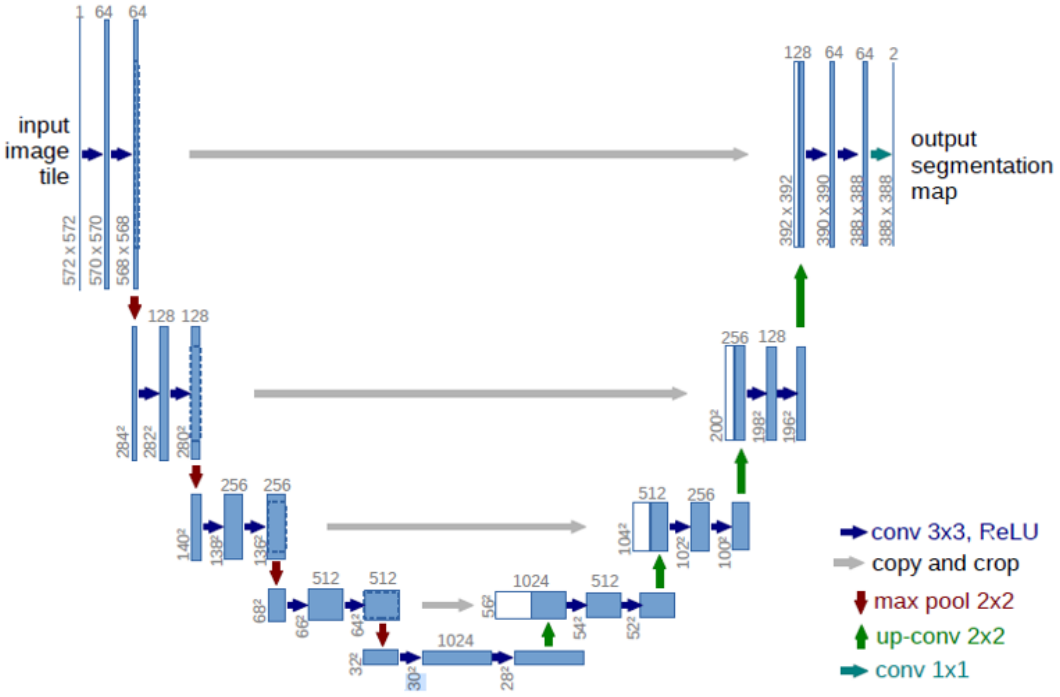

Figure 4.2.1: U-Net architecture. Adapted from 'U-Net: Convolutional Networks for Biomedical Image Segmentation,' by O.Ronneberger, P.Fischer and T.Brox, 2015, International Conference on Medical image computing and computer-assisted intervention, p.234–241.

to transform the channel number from M to desired N, pointwise convolutional filters in (c) are used as 1×1 convolutional filters to transform channel numbers to desirable ones. As described in Howard et al. (2017), depthwise separable convolution can get a reduction in computation of $\frac{1}{N} + \frac{1}{D_k^2}$. Using 3×3 convolutional layer in U-Net as example, this technique leads to around $\frac{1}{8}$ computation cost compared with the classic convolutional filters.

Inverted residual with linear bottleneck Inverted residual with linear bottleneck was proposed in Sandler et al. (2018) and described in Fig 4.2.3. In deep neural network research, there is a notorious degradation problem: with more stacked layers to a deep model, the acuracy becomes saturated and then degrades rapidly. To solve this problem and

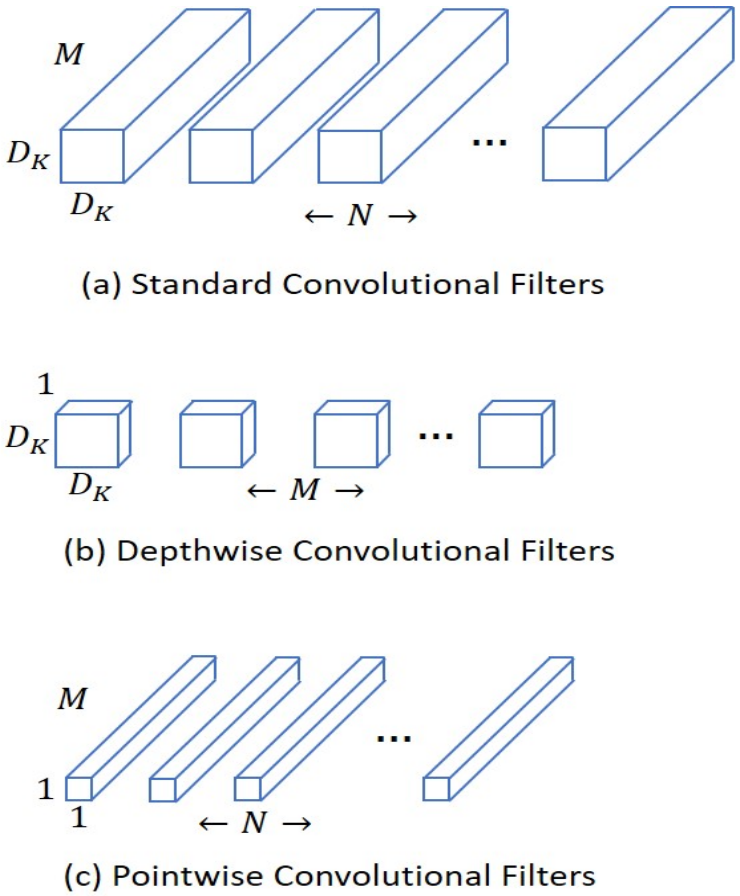

Figure 4.2.2: The classic convolution filters in (a) have been decomposed to depthwise convolution in (b) and pointwise convolution in (c).

create deeper models with promising accuracy, He et al. (2016) proposed ResNet. ResNet includes residual learning blocks to learn residual of desired underlying feature mapping instead of the feature mapping itself, then adds the input feature map to the end of the block. This residual block dramatically relieves the degradation problem that leads to deeper and deeper networks such as ResNet152, which included 152 layers. Inverted residual builds a similar residual block with bottleneck compared with ResNet. The difference is that residual blocks in ResNet are connecting two layers with higher number of channels while inverted residual blocks are connecting two bottleneck layers with low number of channels. Thus, residual blocks have an hourglass-shape while inverted residual blocks are spindle-shaped.

The intuition of inverted residual block is: non-linear function such as ReLU does not work well in low-dimensional space compared with linear functions (Sandler et al., 2018). Instead of connecting two high-dimensional layers, inverted residual blocks are connecting two low-dimensional linear bottleneck layers while the intermediate high-dimensional expansion layers are more efficient to use non-linear activation functions for information retrieval. With this inverted residual block, deep neural networks can be deeper without explosion on number of parameters and relieve the degradation problem.

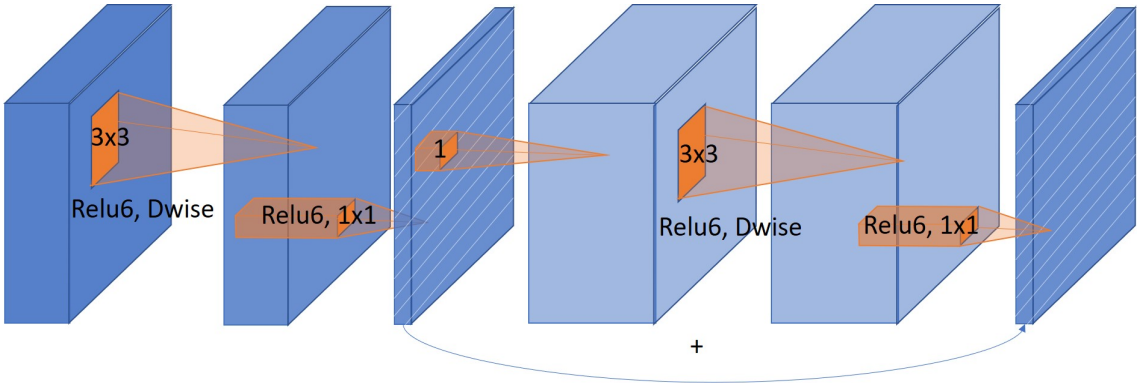

Figure 4.2.3: The inverted residual block inserts a bottle neck layer (diagonally batched layers) between pointwise convolutional layers and output feature map. Then, a inverted residual block is considered as components between two bottleneck layers shown with last 4 layers.

Squeeze-and-Excitation(SE) Squeeze-and-Excitation(SE) was proposed in Hu et al. (2018) and is described in Fig 4.2.4. SE is a powerful tool to build a unit to recalibrate any feature maps. The goal of SE is to selectively emphasize informative features and suppress less useful ones. In Fig 4.2.4, an input feature map X with dimension $H \times W \times C$ is passed to a transformation operation F_{tr} and generates an output feature map U with dimension $H^{'} \times W^{'} \times C^{'}$. Then, a unit built by SE is described in the following steps:

1. U has been squeezed channel-wise by F_{sq}, i.e. calculate the mean of each $H^{'} \times W^{'}$ feature map which resulted in a $1 \times 1 \times C^{'}$ tensor;
2. the squeezed feature map is passed to a self-gating function F_{ex} i.e. a sigmoid activation $s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z))$, where δ refers to the ReLU function,

$W_1 \in \Re^{\frac{C'}{r} \times C'}$ and $W_2 \in C' \times \Re^{\frac{C'}{r}}$. C' is the number of channel of U and r is a prespecified reduction ratio used to build the self-gating mechanism with detailed discussion in Hu et al. (2018);

3. The final output of the block is obtained by rescaling U with the operation F_{scale} i.e. $\tilde{U} = F_{scale}(u_c, s_c) = s_c u_c$, where F_{scale} refers to channel-wise multiplication between the scalar s_c from excitation operation and each channel-wise 2d feature map in U.

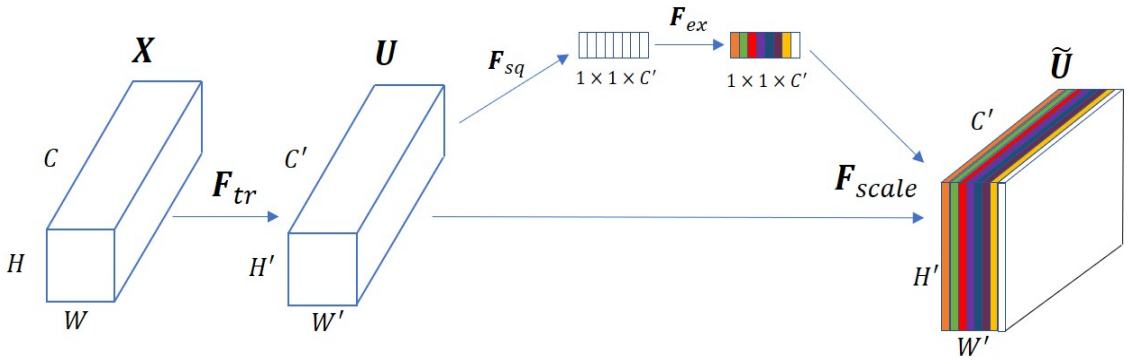

Figure 4.2.4: A Squeeze-and-Excitation block: an output feature map U is first squeezed by a function F_{sq} and followed by an excitation operation with a self-gating function F_{ex}. Output weights from excitation will used to recalibrate U and generate final output feature map $\tilde{U}$ with operation F_{scale}.

ACU-Net convolutional layer block is then built based on above techniques and described in Fig 4.2.5. Fig 4.2.5.(a) shows ACU-Net convolutional layer block without Squeeze-and-Excitation which is the same block built in Sandler et al. (2018). Fig 4.2.5.(b) shows ACU-Net convolutional layer block with Squeeze-and-Excitation which is the same block built in Howard et al. (2019).

4.2.2 ACU-Net architecture

ACU-Net architecture is established based on following two ideas to relieve heavy parameterization problem of U-Net to avoid overfitting. The first idea is *Light-Coder-and-Heavy-Bottleneck* and the second is Asymmetric-Auto-Encoder.

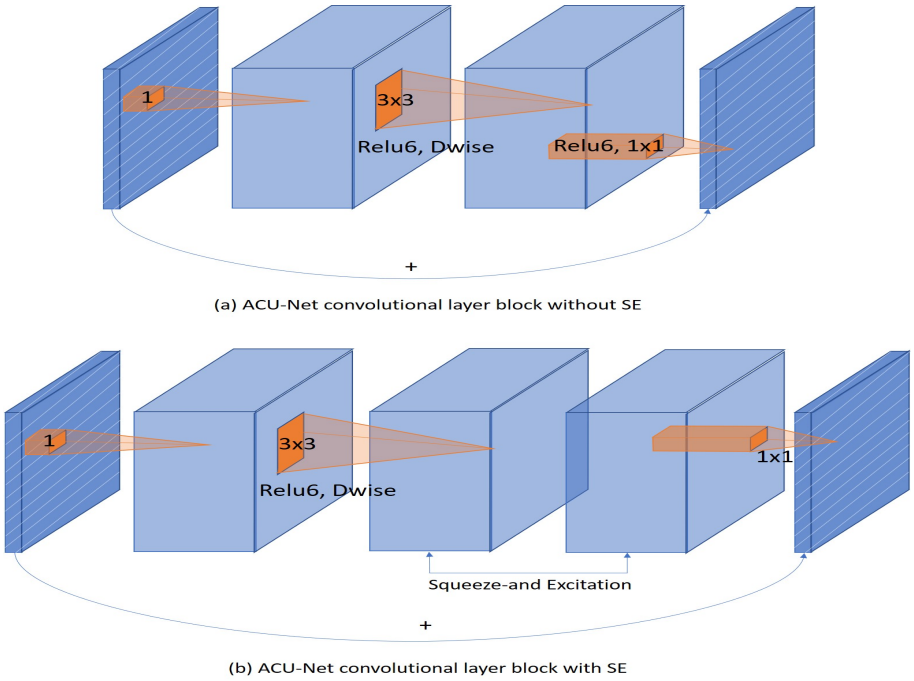

Figure 4.2.5: ACU-Net convolutional layer block without Squeeze-and-Excitation in (a) and with Squeeze-and-Excitation in (b).

Light-Coder-and-Heavy-Bottleneck In original U-Net described in Fig 4.2.1, number of channels of encoders (left part of U-Net) and decoders (right part of U-Net) are doubled in next level layer block. This is a heavy design where the trade off between computation cost and accuracy might not be well-optimized. Ignited by MobileNetV2 in Sandler et al. (2018), low-dimensional bottleneck layer can well preserve the information. Thus, ACU-Net demonstrated in Fig 4.2.6 uses a light encoder and decoder design with much fewer channels compared with U-Net while still keep channel concatenation at the beginning of each decoder block.

Asymmetric-Auto-Encoder In U-Net, each decoder block has the same operation compared with its corresponding encoder block i.e. two convolutional layers operation. Although U-Net has a U-shape symmetric architecture, it is still in sequential order. The double convolutional layer operations in encoder block might be helpful for information retrieval while the corresponding decoder blocks with the same operation might not be able

to keep the same efficacy compared with their encoder counterpart. Thus, decoder parts in ACU-Net in Fig 4.2.6 with green color have fewer operations compared with their corresponding encoder parts with blue color.

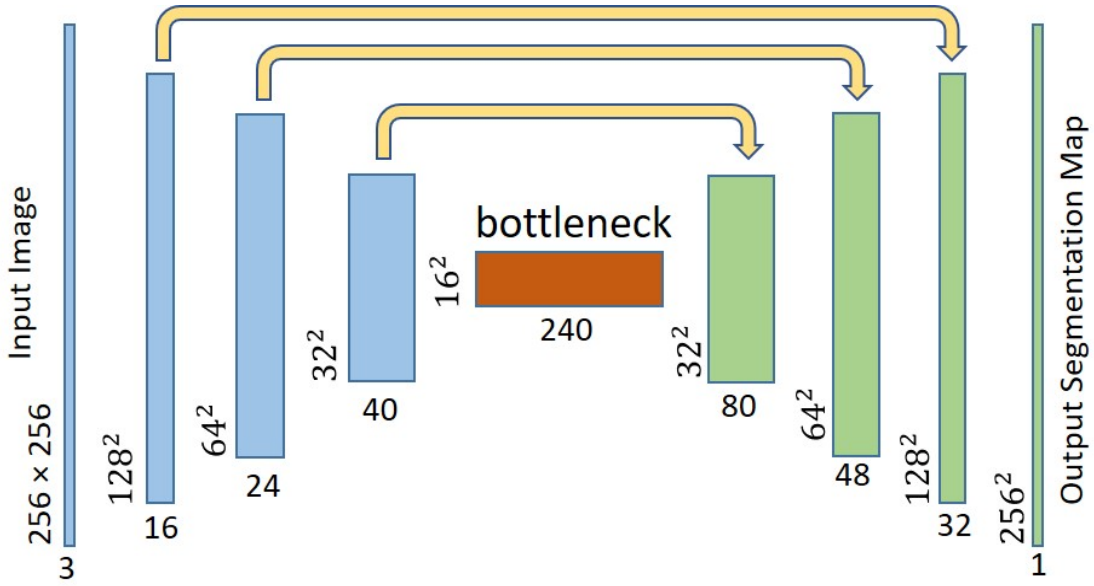

Figure 4.2.6: ACU-Net architecture.

Before introducing the details of components of ACU-Net, we first introduce several definitions that are used in ACU-Net. *Batch Normalization* was proposed in Ioffe and Szegedy (2015), which was used to relieve internal covariate shift i.e. different inputs of each layers slowed down the training by requiring lower learning rates and careful parameter initialization. Batch Normalization (BN) normalizes a part of the model architecture and performing the normalization for each training mini-batch. BN is used in ACU-Net inverted block after each convolutional operation. *Hard swish* activation function is defined as: $h\text{-}swish(x) = x\frac{ReLU6(x+3)}{6}$ where $ReLU6(x) = min(max(x, 0), 6)$ is the clipped version of ReLU. This activation function is well validated in Howard et al. (2019) to avoid gradient vanishing/exploration problem while reduce the number of memory accesses by used in deeper layers of the model. In Table 4.2.1, we list the details of layers in ACU-Net.

Table 4.2.1: Details for ACU-Net

Input	Operator	exp size	#out channel	SE	NL	s
Encoder						
$256^2 \times 3$	conv2d	-	16	-	HS	2
$128^2 \times 16$	InvRes-B, 3×3	16	16	-	RE	1
$128^2 \times 16$	InvRes-B, 3×3	64	24	-	RE	2
$64^2 \times 24$	InvRes-B, 3×3	72	24	-	RE	1
$64^2 \times 24$	InvRes-B, 5×5	96	40	$\surd$	HS	2
$32^2 \times 40$	InvRes-B, 5×5	240	40	$\surd$	HS	1
$32^2 \times 40$	InvRes-B, 5×5	240	40	$\surd$	HS	1
$32^2 \times 40$	InvRes-B, 5×5	240	240	$\surd$	HS	2
Decoder						
$16^2 \times 240$	Upconv2d	-	40	-	-	2
$32^2 \times 40$	Up-InvRes-B, 5×5	240	40	$\surd$	HS	1
$32^2 \times 40$	Upconv2d	-	24	-	-	2
$64^2 \times 24$	Up-InvRes-B, 5×5	72	24	$\surd$	HS	1
$64^2 \times 24$	Upconv2d	-	16	-	-	2
$128^2 \times 24$	Up-InvRes-B, 3×3	16	16	$\surd$	HS	1
$128^2 \times 16$	Upconv2d	-	16	-	-	2
$256^2 \times 16$	conv2d	-	1	-	Sig	1

exp size is expansion layer channel size in ACU-Net convolutional layer block. **InvRes-B**, 3×3 refers to ACU-Net convolutional layer block with 3×3 filter size. **Upconv2d** is up-convolutional layer as same as in U-Net to double the height and width of input feature map while change the number of channels in decoder part. **Up-InvRes-B** is operation which first concatenates encoder part to decoder then followed by InvRes-B operation. **SE** refers to whether uses Squeeze-and-Excitation in a speicific block. **NL** refers to non-linear activation function. **HS** refers use hard-swish acitivation function, **RE** refers to ReLU and **Sig** refers to Sigmoid. s refers to stride.

4.3 Experiments

In this section, we present our experimental results to show the effectiveness of ACU-Net. We report segmentation results on the ongoing normal aging study previously described in first project.

Our models are trained with data augmentation. As described in Shorten and Khoshgoftaar (2019), data augmentation techniques have been widely used and validated in the application to medical image analysis to avoid over-fitting problem of heavy models. For an image object, data augmentation techniques includes: flipping, rotation, shearing, cropping and etc. In Fig 4.3.1, we show an example of data augmentation application to our medical image data. In addition, we use online data augmentation in training models. Compared with offline data augmentation which generates a fixed size of augmented dataset, online data augmentation generates an augmented training dataset in each training iteration step based on different augmentation settings. Thus, online data augmentation can generate infinite training samples if the training iteration number grows. In practice, we include rotation, random horizontal flipping and scaling in our online data augmentation step which can generate augmented data with less heterogeneity.

4.3.1 Normal aging dataset

We use the same data split scheme described in the first project which split the 20 subjects into 15 training subjects and 5 testing subjects. Each subject includes 5 manually tracing slices.

In this dataset, the input FLAIR images have dimensions around 256. Thus, instead of building U-Net, we have built U-Net small, which just halve the dimensions of initial input images and resulted in halving dimensions of all following feature maps step by step.

Training setup We trained our models on a 8GB GTX 1080 GPU. We use the standard Adam optimizer (Kingma and Ba, 2014) with initial learning rate of 0.01. The mini-batch size is set to 15. We use dropout (Srivastava et al., 2014) with rate as 0.5 to last output layer.

Measurement setup Since we only have 25 testing slices in this dataset, we have created an augmented testing dataset includes 500 augmented images from the original 25 testing slices. The performance metrics we use are same as metrics used in the first project. In addition, we use number of parameters and FLOPs to measure the efficiency of models. FLOPs is the floating point operations which measures the complexity of the model.

Results The performance comparison is described in Table 4.3.1. We can find ACU-Net only loses 2% DSC on original testing dataset and 1% DSC augmented testing dataste while achieves around 1/20 model size and 1/40 complexity compared with U-Net-small.

4.4 Conclusions

The proposed ACU-Net represents a novel compact convolutional neural network based on a well-validated architecture U-Net. The goal of ACU-Net is to build an efficient compact convolutional neural network for biomedical image segmentation. Thus, ACU-Net builds an inverted residual block with linear bottleneck and squeeze-and-excitation for convolutional layers block. In addition, ACU-Net builds a new asymmetric auto-encoder architecture with more weights on encoders part. This architecture decreases computation cost on decoders part while preserves the model performance. Compared with U-Net, ACU-Net focuses more on the information passing to bottleneck layer in the full architecture, thus, ACU-Net decreases the number of channels used in encoders and decoders part while keeps the high channel numbers in the bottom bottleneck layer. ACU-Net achieves competitive model performance compared with U-Net on a normal aging cohort WMH segmentation problem while decreases the model size and model complexity to 1/20 and 1/40 of U-Net respectively. This efficient structure of ACU-Net is favorable since modern CNNs require more and more computation resources while in many research environments, the computation resources are limiting. ACU-Net's compact model size enables researchers to train the model from scratch with their own data instead of using pre-trained models due to limited computation resources. It is even possible to move ACU-Net to mobile devices in the future since its convolutional layers block are based on blocks built in MobileNets which are designed for mobile devices.

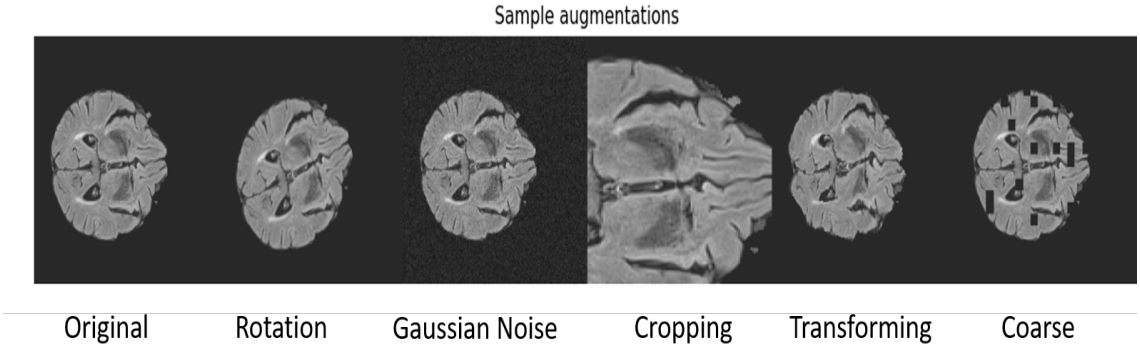

Figure 4.3.1: Data augmentation case study example

Table 4.3.1: Performance comparison

	ACC	PPV	TPR	FPR	DSC	AUC	Params(M)	FLOPs(G)
Original testing dataset (25 slices)								
U-Net-small	0.985(0.003)	0.876(0.037)	0.767(0.068)	0.005(0.001)	0.817(0.049)	0.881(0.034)	7.76	13.72
ACU-Net	0.985(0.002)	0.866(0.055)	0.753(0.092)	0.005(0.002)	0.801(0.06)	0.874(0.045)	**0.37**	**0.39**
Augmented testing dataset (500 augmented slices)								
U-Net-small	0.984(0.007)	0.856(0.062)	0.773(0.08)	0.006(0.002)	0.809(0.059)	0.953(0.023)		
ACU-Net	0.983(0.007)	0.861(0.066)	0.758(0.097)	0.006(0.002)	0.801(0.063)	0.884(0.056)		

5.0 Discussions

Segmentation of WMHs is an important task with biological meanings. Accurate automatic segmentation of WMHs not only help physicians save time on manual tracing of neuroimages of patients, but also ensure the accuracy of automatic tissue segmentation of the brain by filling out WMHs with normal white matter tissues such that those WMH regions will not be wrongly classified as grey matter tissues. Thus, in our first project, we propose an improved automatic voxel-wise WMH segmentation tool OASIS-AD based on logistic regression that can handle a small sample size of neuroimages, which is a very common scenario in medical studies. OASIS-AD uses two refined method, NNR and GFR, in combination to reduce the false positive rate of WMHs segmentation, especially due to speckling. In particular, NNR uses neighborhood information combined with information from the `FAST` segmentation algorithm to increase or decrease the estimated probability that a voxel should be identified as WMH. A potential disadvantage of NNR and GFR is that in certain situations they may lead to results that are too conservative when probabilities are shrunk too aggressively towards zero. There are several potential solutions that could be considered to help address these problems. For example, in the first step of the NNR(v) algorithm described in Section 1 we used the transformation $P^{rv}_{wmh} = (P^{v}_{wmh})^{10}$ for voxels that were estimated by `FAST` to have probability 1 of being in white matter and all 6NN to be in white matter. One could use alternative transformations and one could better use `FAST`, or other segmentation algorithms, to inform the likelihood that the voxel is in WMH. One solution could be to use `FAST` and OASIS-AD iteratively: first use `FAST` to segment white matter, gray matter, and CSF and then use OASIS-AD to estimate the WMH. Once this is done the WMH region estimated via OASIS-AD can be filled in with normally appearing white matter and the process could be iterated until no differences are observed. In our study we only have two image modalities, T1 and FLAIR, though OASIS-AD can be easily extended to incorporate additional image modalities, while standard variable selection techniques, as well as interaction terms, could easily be embedded in the model structure.

In our second project, we develop a high-dimensional image-on-scalar regression model

IRRR to reveal the association between neuro-degenerative level and patients' cardiovascular predictors. As described in Section 3.7, this method is not exhaustive and suggests several future research directions, including methodological, theoretical and numerical extensions. Methodologically, future research will investigate the potential extension to nonlinear dimension reduction. The estimators in IRRR procedure can be modified to the setting where one does not have a selected group of predictors but desires to use a large number of predictors, only a subset of which can be associated with the image response variables. This can be done by combining the lasso-based estimator of B proposed by Ma and Sun (2014) with the proposed estimator of V in Equation 3.3 in a two-step algorithm. Theoretically, besides the future extension on the consistency results described in Section 3.7, the noise setting is also a potential extension directions. Although we assume Gaussian noise, we conjecture that the statistical properties could hold for more general, non-Gaussian distributions. In addition, consistency proofs were established, but not oracle properties for model selection. We hypothesize that the incorporation of adaptive weights (Beer et al., 2019) could lead to consistent model selection. Numerically, although ADMM is a convenient optimization tool for complicated regularization structures, and we constructed an efficient package for implementing the proposed method, there exist other sophisticated optimization methods that could potentially be used. Future research could include the investigation of other numerical methods and a formal analysis of their computational costs relative to the proposed ADMM.

As an extension of first two projects, our third project developed a deep neural network that inspired by the decomposition techniques used in second project to solve the WMHs segmentation problem with a relatively small sample size of images. Deep learning approaches can provide an alternative to OASIS-AD and we continue to investigate the added benefit of these techniques, including convolutional neural networks. So far, we have seen encouraging results, though much remains to be done in terms of increasing the sample size of the training data (not easy to achieve in low resource environments), performance (we have not yet matched OASIS-AD), interpretability (we would like to better understand what features of the data are actually contributing to improved prediction performance), and choices of the many tuning parameters (e.g., neighborhood size and filter types). Thus, it is still very challenging to adopt modern DNNs to medical image segmentation studies. Among those DNNs,

U-Net might be the most doable DNN segmentation tool that can work with a small sample size of images. Based on U-Net, we built ACU-Net with modern compression techniques. ACU-Net can dramatically decrease the number of parameters and model complexity compared with U-Net while keep the similar model performance on WMH segmentation tasks. In original U-Net, 3×3 filters are used in convolutional layers, however, with decomposition of filters, we can use larger filters with even less parameters. Besides the benefits of ACU-Net discussed in Section 4.4, a potential extension of ACU-Net is to follow ResNet152 to stack layers in the networks. Although more layers are added to the model which cost more computation cost, the compact design of convolutional layers block can efficiently to improve the model performance while still keep lower computation cost compared with classic CNN architecture.

In conclusion, in the future time, I will work on: (1) publish the work in the second and third projects. (2) apply ACU-Net to more open source data challenges to test its performance (3) release the R-packages for the second project to the public and release the Python library of the third project to the public.

Appendix A

Supplementary materials for Chapter 2

A.1 Extra Table

Table A.1.1: Performance evaluation metrics(full)

	ACC	PPV	TPR	FPR	DSC	ROC	PRC
OASIS-AD(M1)	0.96(0.01)	0.84(0.11)	0.58(0.07)	0.008(0.005)	0.69(0.07)	0.95	0.80
OASIS-AD(M1-N)	0.96(0.01)	0.86(0.11)	0.5(0.1)	0.006(0.004)	0.63(0.1)	0.83	0.70
OASIS-AD(M1-G)	0.97(0.01)	0.85(0.08)	0.69(0.07)	0.009(0.003)	0.79(0.06)	0.97	0.86
OASIS-AD(M1-NG)	0.96(0.01)	0.86(0.09)	0.63(0.11)	0.007(0.004)	0.72(0.09)	0.95	0.82
OASIS-AD(M1-GN)	0.96(0.01)	0.85(0.09)	0.58(0.13)	0.007(0.003)	0.68(0.12)	0.82	0.73
OASIS-AD(M2)	0.95(0.01)	0.8(0.15)	0.52(0.08)	0.009(0.006)	0.63(0.1)	0.94	0.76
OASIS-AD(M2-N)	0.95(0.01)	0.79(0.17)	0.47(0.13)	0.008(0.006)	0.58(0.15)	0.81	0.65
OASIS-AD(M2-G)	0.97(0.01)	0.85(0.11)	0.65(0.1)	0.009(0.007)	0.73(0.08)	0.97	0.86
OASIS-AD(M2-NG)	0.96(0.01)	0.85(0.12)	0.57(0.15)	0.007(0.007)	0.67(0.14)	0.94	0.81
OASIS-AD(M2-GN)	0.96(0.01)	0.84(0.12)	0.54(0.16)	0.007(0.006)	0.65(0.16)	0.82	0.72
OASIS	0.95(0.02)	0.76(0.11)	0.59(0.11)	0.014(0.004)	0.66(0.11)	0.92	0.74
MIMOSA	0.97(0.01)	0.94(0.06)	0.58(0.11)	0.002(0.001)	0.72(0.1)	0.87	0.77
LST	0.97(0.01)	0.84(0.13)	0.72(0.12)	0.012(0.012)	0.76(0.07)	0.87	0.77
fuzzy-c	0.95(0.01)	0.90(0.12)	0.49(0.13)	0.019(0.014)	0.63(0.11)	NA	NA

A.2 Extra Figure

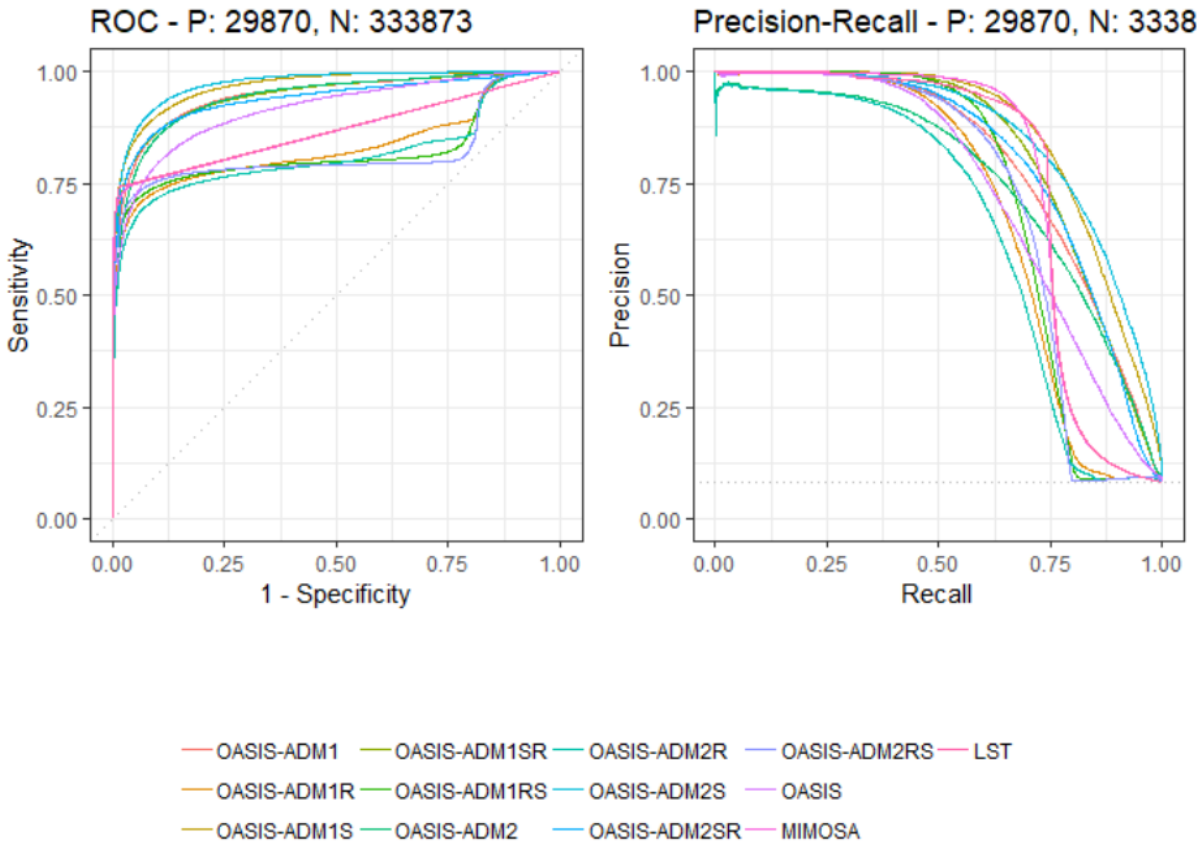

Figure A.2.1: ROC and PRC of models(full)

Appendix B

Supplementary materials for Chapter 3

B.1 Extra Figures

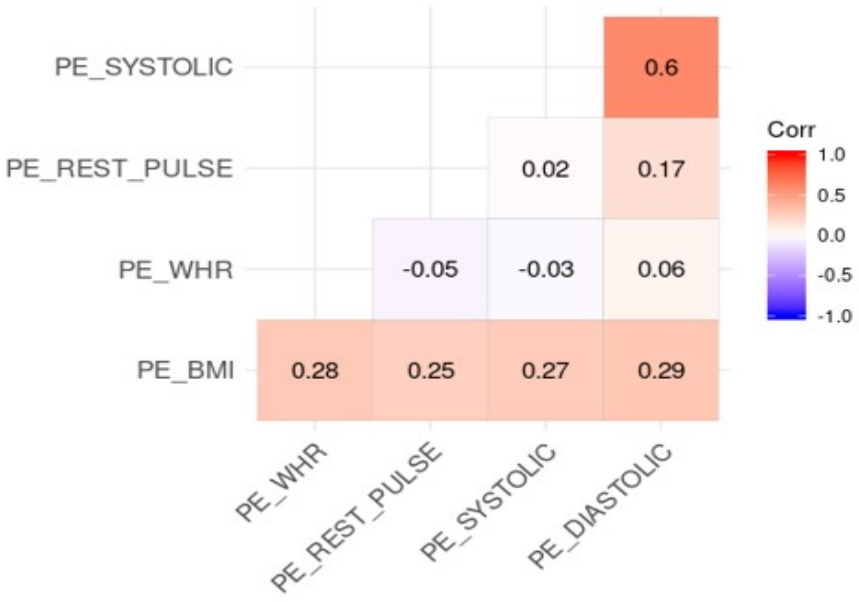

Figure B.1.1: Correlation plot of predictors

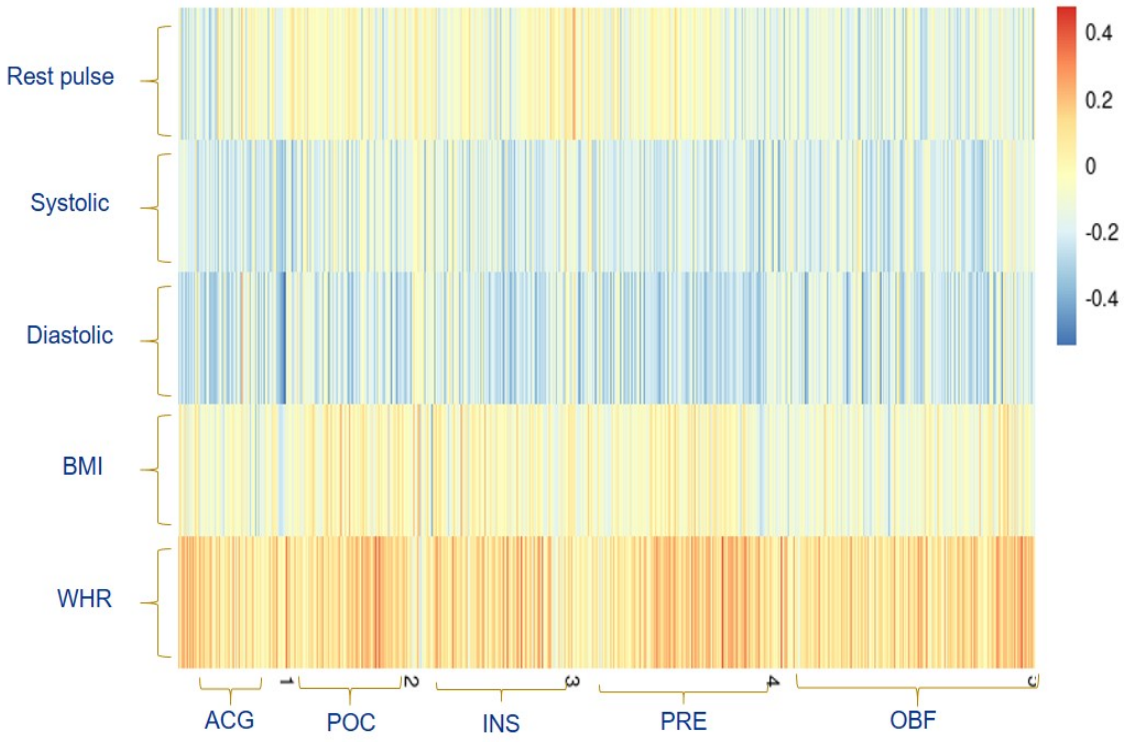

Figure B.1.2: Univariate correlation analysis between predictors and voxels

B.2 Technical Details and Illustrations

B.2.1 Fused lasso generalized coefficient matrix D

For a 2D image of dimension $m_1 \times m_2$, so that $m = m_1 m_2$, the fused lasso for each predictor is the sum of $m_F = 2m - m_1 - m_2$ absolute differences between adjacent pixels. For example, for a 2×2 image, there are $m_F = 4$ terms and the matrix $D \in \Re^{m_F \times m} = \Re^{4\times 4}$ for the generalized lasso representation can be expressed as

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

For a 3D image of dimension $m_1 \times m_2 \times m_3$, so that $m = m_1 m_2 m_3$, the fused lasso for each predictor is the sum of $m_F = (3m - m_1 m_2 - m_1 m_3 - m_2 m_3)$ absolute differences between adjacent voxels. For example, for a $2 \times 2 \times 2$ image, there are $m_F = 12$ absolute differences and the matrix $D \in \Re^{m_F \times m} = \Re^{12\times 8}$ for the generalized lasso representation can be expressed as

$$D = \begin{bmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

B.2.2 Notations and assumptions

We first introduce the notations of sparsity condition of penalty terms with a given matrix B^* defined in Theorem 1. Let $J_1 = \{ij : |B_i^* V_j^T| \neq 0\}$, $1 \leq i \leq p, 1 \leq j \leq m$ be the index set of nonzero elements in sparse lasso penalty terms, where B_i^* is ith row vector of B^* and V_j is jth row vector of V; let $J_2 = \{ij : |B_i^*(DV)_j^T| \neq 0\}$, $1 \leq i \leq p, 1 \leq j \leq m_F$ be the index set of nonzero elements in fused penalty terms, where D is fused lasso generalized

coefficient matrix defined in last section and let $(DV)_j$ is jth row vector of DV; $J_3 = \{g \in \mathbb{G} : ||B^*(GV)_g^T||_2 \neq 0\}$ be the index set of nonzero elements in group penalty terms, where $\mathbb{G}$ is a set introduced in Section 3.2. In addition, let J_1^c, J_2^c and J_3^c be the complementary index set of J_1, J_2 and J_3 respectively. Let $s_1 = |J_1|$, $s_2 = |J_2|$ and $s_3 = |J_3|$ be the number of elements in each index set, where s_1, s_2 and s_3 are finite numbers.

Next, we introduce notations of projection: for any nontrivial matrix $\Delta \in \Re^{m\times r}$, let $B^*\Delta_{J_1}^T$ represent the projection of $B^*\Delta^T$ on J_1 i.e. $B^*\Delta_{J_1}^T$ is sub-matrix of $B^*\Delta^T$ whose columns are in set J_1 columns of $B^*\Delta^T$; $B^*\Delta^T D_{J_2}^T$ represent the projection of $B^*\Delta^T D^T$ on J_2; $B^*\Delta^T G_{J_3}^T$ represent the projection of $B^*\Delta^T G^T$ on J_3, and apply the same notations rules for all complementary sets and elements in any sets.

Lastly, we introduce measure of sparsity S used in Theorem 1. Let $\lambda_g = \lambda_3 m_g^{1/2}$. we define

$$S = \frac{s_1^{1/2}}{k_1 n^{1/2}} + \frac{s_2^{1/2}(\lambda_2/\lambda_1)\sigma_{D^+}}{k_2 n^{1/2}} + \frac{\left[\sum_{g\in J_3}(\lambda_g/\lambda_1)^2\right]^{1/2}}{k_3 n^{1/2}},$$

where σ_{D^+} is the largest singular value of D, which is proved to be finite in the following Lemma 2, and k_1, k_2, k_3 are defined in Assumption 1.

Assumption 1. *Let $J_1 \subseteq \{ij : 1 \leq i \leq p, 1 \leq j \leq m\}$, $J_2 \subseteq \{ij : 1 \leq i \leq p, 1 \leq j \leq m_F\}$ and $J_3 \subseteq \{1, \ldots, |\mathbb{G}|\}$ be any index sets s.t. $J_1 \leq s_1$, $J_2 \leq s_2$ and $J_3 \leq s_3$. Let β be a positive number and $\zeta = \{\zeta_g : g \in \mathbb{G}\}$ be a set of positive numbers. D is the generalized fused lasso coefficient matrix in a 3-dimensional space. For a given matrix $B^* \in \Re^{p\times r}$, and any nontrivial matrix $\Delta \in \Re^{m\times r}$ that satisfies*

$$\begin{aligned}
&|B^*\Delta_{J_1^c}^T|_1 + 2\beta|B^*\Delta^T D_{J_2^c}^T|_1 + 2\sum_{g\in J_3^c}\zeta_g||B^*\Delta^T G_g^T||_2 \\
&\leq 3|B^*\Delta_{J_1}^T|_1 + 2\beta|B^*\Delta^T D_{J_2}^T|_1 + 2\sum_{g\in J_3}\zeta_g||B^*\Delta^T G_g^T||_2,
\end{aligned}$$

the following minimums exist and are positive:

$$\begin{aligned}
k_1 &= \min_{J_1,J_2,J_3,\Delta\neq 0} \frac{||XB^*\Delta^T||_2}{n^{1/2}||B^*\Delta_{J_1}^T||_2}, \\
k_2 &= \min_{J_1,J_2,J_3,\Delta\neq 0} \frac{||XB^*\Delta^T D^T||_2}{n^{1/2}||B^*\Delta^T D_{J_2}^T||_2}, \\
k_3 &= \min_{J_1,J_2,J_3,\Delta\neq 0} \frac{||XB^*\Delta^T||_2}{n^{1/2}||B^*\Delta^T G_{J_3}^T||_2}.
\end{aligned}$$

Assumption 1 restricts non-signal terms by signal terms of any nontrivial matrix. In addition, it assumes existence of positive minimums of singular values of projection matrix of sparse lasso, fused lasso and group lasso penalty matrix.

Assumption 2. *For the coefficient matrix A in (3.1), A belongs to the following parameter space:*

$$\Theta(r, d, \tau) = \{A \in \Re^{p\times m} : rank(A) = r, \tau d \geq \sigma_1(A) \geq ... \geq \sigma_r(A) > d > 0\}.$$

In addition, after rank factorization of A i.e. $A = BV^T$, with out loss of generality, we assume $||V||_2 = 1$.

Assumption 3. *Let $\hat{B}$ be an estimator of B^*, which is true left singular space of A^* s.t. $A^* = B^*V^{*T}$. We assume*

$$\hat{B} = B^* + O(1/\sqrt{n}).$$

Assumption 2 restricts A to a space with finite rank and finite singular values. Assumption 3 restricts $\hat{B}$ to be a $\sqrt{n}$-consistent estimator. These two assumptions are established for the fixed p large m and n setting as described in Section **??**.

B.2.3 Additional lemmas

Lemma 1. *Let $X_{B^*} = XB^*$, $\lambda_1 = 2Cd_X\sigma\sqrt{log(pm)}$, where C is a constant s.t. $C > \sqrt{2}$. Let B_i^* be column vectors of B^* where $i = 1, \ldots, p$. Let $\hat{V}$ be the minimizer of (3.3). With probability at least $1 - (pm)^{1-C^2/2}$, we have*

$$\begin{aligned}
&\frac{1}{2}||X_{B^*}(V^* - \hat{V})^T||_2^2 + \lambda_1|B^*(\hat{V} - V)^T|_1 + 2\lambda_2|B^*(\hat{V} - V)^T D^T|_1 + 2\sum_{g\in\mathbb{G}} \lambda_g||B^*(\hat{V} - V)^T G_g^T||_2 \\
&\leq \frac{1}{2}||X_{B^*}(V^* - V)^T||_2^2 + 4\lambda_1 \sum_{ij\in J_1} |B_i^*(\hat{V} - V)_j^T| \\
&\quad + 4\lambda_2 \sum_{ij\in J_2} |B_i^*(\hat{V} - V)^T D_j^T| + 4\sum_{g\in J_3} \lambda_g||B^*(\hat{V} - V)^T G_g^T||_2. \qquad \text{(B.1)}
\end{aligned}$$

Lemma 2. *Let D be a generalized lasso coefficient matrix of fused lasso structure in 3-dimensional space. Let σ_D^+ be the largest singular values of D, we have*

$$\sigma_D^+ \leq 2\sqrt{3}$$

B.2.4 Proof of Lemma 1

Proof. Since $\hat{V}$ is minimizer of (3.3), thus, for any $V \in \Re^{m\times r}$, we have

$$\frac{1}{2}||Y - X_{B^*}\hat{V}^T||_2^2 + 2\lambda_1|B^*\hat{V}^T|_1 + 2\lambda_2|B^*\hat{V}^T D^T|_1 + 2\sum_{g\in\mathbb{G}}\lambda_g||B^*\hat{V}^T G_g^T||_2$$
$$\leq \frac{1}{2}||Y - X_{B^*}V||_2^2 + 2\lambda_1|B^*V^T|_1 + 2\lambda_2|B^*V^T D^T|_1 + 2\sum_{g\in\mathbb{G}}\lambda_g||B^*V^T G_g^T||_2.$$

Plugging $Y = X_{B^*}V^{*T} + E$ into above inequality, we have

$$\begin{aligned}\frac{1}{2}||X_{B^*}(V^* - \hat{V})^T||_2^2 \leq \frac{1}{2}||X_{B^*}(V^* - V)^T||_2^2 &+ \sum_{k=1}^{n}\sum_{j=1}^{m}[X_{B^*}(\hat{V} - V)^T]_{kj}e_{kj} \\ &+ 2\lambda_1(|B^*V^T|_1 - |B^*\hat{V}^T|_1) \\ &+ 2\lambda_2(|B^*V^T D^T|_1 - |B^*\hat{V}^T D^T|_1) \\ &+ 2\sum_{g\in\mathbb{G}}\lambda_g(||B^*V^T G_g^T||_2 - ||B^*\hat{V}^T G_g^T||_2).\end{aligned}$$

where $[X_{B^*}(\hat{V}-V)^T]_{kj}$ is the $(kj)^{th}$ elements of matrix $X_{B^*}(\hat{V}-V)^T$ and e_{kj} is $(kj)^{th}$ element of matrix E, $1 \leq k \leq n, 1 \leq j \leq m$. For these error related terms, we have

$$\sum_{k=1}^{n}\sum_{j=1}^{m}[X_{B^*}(\hat{V} - V)^T]_{kj}e_{kj} = \sum_{k=1}^{n}\left\{\sum_{j=1}^{m}\left[\sum_{i=1}^{p}X_{ki}(B_i^*\hat{V}_j^T - B_i^*V_j^T)\right]e_{kj}\right\}$$
$$\leq \max_{1\leq i\leq p, 1\leq j\leq m}\left|\sum_{k=1}^{n}X_{ki}e_{kj}\right|\sum_{i=1}^{p}\sum_{j=1}^{m}|B_i^*\hat{V}_j^T - B_i^*V_j^T| = |X^TE|_\infty|B^*\hat{V}^T - B^*V^T|_1,$$

where $|X^TE|_\infty$ is infinity norm of X^TE.

Now, let $\omega_{ij} = X_i^Te_j$, $1 \leq i \leq p, 1 \leq j \leq m$. Let d_i be the ith diagonal element of X^TX/n. Since $e_j \sim N(0, \sigma I_n)$, it is trivial $\text{var}(\omega_{ij}) = X_i^T\text{cov}(e_j)X_i = nd_i^2\sigma^2$. Thus, $(nd_i^2\sigma^2)^{-1/2}\omega_{ij}$ are standard normal random variables. Consider following random event of ω_{ij}

$$\mathbb{C} = \left\{|X^TE|_\infty \leq \lambda_1\right\},$$

and its complementary event is

$$\mathbb{C}^c = \left\{\text{at least one } |\omega_{ij}| > \lambda_1, 1 \leq i \leq p, 1 \leq j \leq m\right\}.$$

Let $U\left(0, \lambda_1\right)$ be a 1-dimensional space centered at 0 with radius λ_1, then we have

$$
\begin{aligned}
Pr\{\mathbb{C}^c\} &\leq \sum_{i=1}^{p}\sum_{j=1}^{m} Pr\left\{\omega_{ij} \notin U\left(0, \lambda_1\right)\right\} = \sum_{i=1}^{p}\sum_{j=1}^{m} Pr\left\{(nd_i^2\sigma^2)^{-1/2}\omega_{ij} \notin U\left(0, \frac{\lambda_1}{2d_i\sigma n^{1/2}}\right)\right\} \\
&\leq \sum_{i=1}^{p}\sum_{j=1}^{m} Pr\left\{|Z| \geq \frac{\lambda_1}{2d_i\sigma n^{1/2}}\right\} \leq \sum_{i=1}^{p}\sum_{j=1}^{m} \exp\left(\frac{-\lambda_1^2}{8nd_i^2\sigma^2}\right) \leq pm\, \exp\left(\frac{-\lambda_1^2}{8nd_X^2\sigma^2}\right) \\
&= (pm)^{1-C^2/2},
\end{aligned}
$$

where Z represents standard normal random variable and d_X is the maximum diagonal element of X^TX/n described in Section 3.4. Last inequality is obtained by using trivial tail bound property of Z which is $Pr\{|Z| > \alpha\} \leq \exp(-\alpha^2/2)$, where α is any real number here. Then, on event $\mathbb{C}$, we have

$$
\begin{aligned}
&\frac{1}{2}||X_{B^*}(V^* - \hat{V})^T||_2^2 + \lambda_1|B^*(\hat{V} - V)^T|_1 + 2\lambda_2|B^*(\hat{V} - V)^TD^T|_1 + 2\sum_{g\in\mathbb{G}}\lambda_g||B^*(\hat{V} - V)^TG_g^T||_2 \\
&\leq \frac{1}{2}||X_{B^*}(V^* - V)^T||_2^2 + \sum_{k=1}^{n}\sum_{j=1}^{m}[X_{B^*}(\hat{V} - V)^T]_{kj}e_{kj} + \lambda_1|B^*(\hat{V} - V)^T|_1 \\
&\quad + 2\lambda_1(|B^*V^T|_1 - |B^*\hat{V}^T|_1) + 2\lambda_2(|B^*(\hat{V} - V)^TD^T|_1 + |B^*V^TD^T|_1 - |B^*\hat{V}^TD^T|_1) \\
&\quad + 2\sum_{g\in\mathbb{G}}(\lambda_g||B^*(\hat{V} - V)^TG_g^T||_2 + ||B^*V^TG_g^T||_2 - ||B^*\hat{V}^TG_g^T||_2) \\
&\leq \frac{1}{2}||X_{B^*}(V^* - V)^T||_2^2 + 2\lambda_1(|B^*(\hat{V} - V)^T|_1 + |B^*V^T|_1 - |B^*\hat{V}^T|_1) \\
&\quad + 2\lambda_2(|B^*(\hat{V} - V)^TD^T|_1 + |B^*V^TD^T|_1 - |B^*\hat{V}^TD^T|_1) \\
&\quad + 2\sum_{g\in\mathbb{G}}\lambda_g(||B^*(\hat{V} - V)^TG_g^T||_2 + ||B^*V^TG_g^T||_2 - ||B^*\hat{V}^TG_g^T||_2) \\
&\leq \frac{1}{2}||X_{B^*}(V^* - V)^T||_2^2 + 4\lambda_1\sum_{ij\in J_1}|B_i^*(\hat{V}_j - V_j)^T| \\
&\quad + 4\lambda_2\sum_{ij\in J_2}|B_i^*(\hat{V} - V)^TD_j^T| + 4\sum_{g\in J_3}\lambda_g||B^*(\hat{V} - V)^TG_g^T||_2.
\end{aligned}
$$

Last inequality is obtained in following way, using $|B^*(\hat{V} - V)^TD^T|_1 + |B^*V^TD^T|_1 - |B^*\hat{V}^TD^T|_1$ as example. $|B^*(\hat{V}-V)^TD^T|_1$ can be split into signal and non-signal parts, where non-signal parts with index set J_1^c are $B^*V = 0$. Thus, $|B^*(\hat{V} - V)^TD^T|_1 + |B^*V^TD^T|_1 - |B^*\hat{V}^TD^T|_1$ on J_1^c is $|B^*\hat{V}^TD^T|_1 - |B^*\hat{V}^TD^T|_1 = 0$. Then, for the signal part, we have

$$
\left(|B^*(\hat{V} - V)^TD^T|_1 + |B^*V^TD^T|_1 - |B^*\hat{V}^TD^T|_1\right)_{J_1} \leq 2\left(|B^*(\hat{V} - V)^TD^T|_1\right)_{J_1}.
$$

The same calculation is applied on fused lasso and group lasso penalty terms, which leads to last inequality. This completes the proof of the lemma. □

B.2.5 Proof of Lemma 2

Proof. Nonzero elements in matrix D are defined by pair of adjacent neighbors. In a 3-dimensional space, for each coordinate, it has up to 6 adjacent neighbors, thus $||D||_\infty = 6$. Since each row of D_{3d} is always with two non-zero elements which are 1 and -1 respectively, and others are 0, thus $||D_{3d}||_1 = 2$. From Horn and Johnson (1991), we have

$$\sigma_D^+ \leq \left(||D_{3d}||_1 ||D_{3d}||_\infty \right)^{1/2} = 2\sqrt{3}.$$

This completes proof of the lemma. □

Lemma 1 shows with probability at least $1 - (pm)^{1-C^2/2}$, error bound of $\hat{V}$ can be restricted by signal's error bound. Lemma 2 shows D's singular values are always bounded by a constant which is not related to dimensions of D. These two lemmas are used in following proof of Theorem 1.

B.2.6 Proof of Theorem 1

Proof. By setting $V = V^*$ in (B.1) in Lemma 1, on event $\mathbb{C}$, we have

$$\begin{aligned}
&\frac{1}{2}||X_{B^*}(\hat{V} - V^*)^T||_2^2 \qquad \text{(B.2)}\\
&\leq 4\lambda_1 \sum_{ij \in J_1} |B_i^*(\hat{V}_j - V_j^*)^T|_1 + 4\lambda_2 \sum_{ij \in J_2} |B_i^*(\hat{V} - V^*)^T D_j^T|_1 + 4 \sum_{g \in J_3} \lambda_g ||B^*(\hat{V} - V^*)^T G_j^T||_2 \\
&\leq 4\lambda_1 s_1^{1/2} ||B_i^*(\hat{V}_j - V_j^*)_{J_1}^T||_2 + 4\lambda_2 s_2^{1/2} ||B^*(\hat{V} - V^*)^T D_{J_2}^T||_2 + 4\Big(\sum_{g \in J_3} \lambda_g^2 \Big)^{1/2} ||B^*(\hat{V} - V^*)^T G_{J_3}^T||_2.
\end{aligned}$$

Last inequality is obtained by Cauchy-Schwarz inequality.

Also by setting $V = V^*$ in inequality (B.1), on event $\mathbb{C}$, we have

$$\begin{aligned}
&\lambda_1 |B^*(\hat{V} - V^*)^T|_1 + 2\lambda_2 |B^*(\hat{V} - V^*)^T D^T|_1 + 2 \sum_{g \in \mathbb{G}} \lambda_g ||B^*(\hat{V} - V^*)^T G_g^T||_2 \\
&\leq 4\lambda_1 \sum_{ij \in J_1} |B_i^*(\hat{V}_j - V_j^*)^T|_1 + 4\lambda_2 \sum_{ij \in J_2} |B^*(\hat{V} - V^*)^T D^T|_1 + 4 \sum_{g \in J_3} \lambda_g ||B^*(\hat{V} - V^*)^T G_g^T||_2.
\end{aligned}$$

By splitting left part of last inequality into signal and non-signal part, we have

$$\lambda_1 \sum_{ij \in J_1^c} |B_i^*(\hat{V}_j - V_j^*)^T|_1 + 2\lambda_2 \sum_{ij \in J_2^c} |B_i^*(\hat{V} - V^*)^T D_j^T|_1 + 2 \sum_{g \in J_3^c} \lambda_g ||B^*(\hat{V} - V^*)^T G_g^T||_2$$

$$\leq 3\lambda_1 \sum_{ij\in J_1} |B_i^*(\hat{V}_j - V_j^*)^T|_1 + 2\lambda_2 \sum_{ij\in J_2} |B_i^*(\hat{V}_f - V_f^*)^T|_1 + 2\sum_{g\in J_3} \lambda_g ||B^*(\hat{V} - V^*)^T G_g^T||_2.$$

Thus, when the condition in Assumption 1 holds, let $\Delta = \hat{V} - V^*$, $\beta = \lambda_2/\lambda_1$ and $\zeta_g = \lambda_g/\lambda_1$, we have

$$\begin{aligned}
||B^*(\hat{V} - V^*)_{J_1}^T||_2 &\leq \frac{||XB^*(\hat{V} - V^*)^T||_2}{k_1 n^{1/2}} \\
||B^*(\hat{V} - V^*)^T D_{J_2}^T||_2 &\leq \frac{||XB^*(\hat{V} - V^*)^T D^T||_2}{k_2 n^{1/2}} \\
&\leq \frac{\sigma_{D^+}||XB^*(\hat{V} - V^*)^T||_2}{k_2 n^{1/2}} \\
||B^*(\hat{V} - V^*)^T G_{J_3}^T||_2 &\leq \frac{||XB^*(\hat{V} - V^*)^T||_2}{k_3 n^{1/2}},
\end{aligned}$$

where σ_{D^+} is defined in Lemma 2. Then, plug the above three inequalities into (B.2), we have

$$\begin{aligned}
&\frac{1}{2}||X_{B^*}(\hat{V} - V^*)^T||_2^2 \\
&\leq \left(\frac{4\lambda_1 s_1^{1/2}}{k_1 n^{1/2}} + \frac{4\lambda_2 s_2^{1/2}\sigma_{D^+}}{k_2 n^{1/2}} + \frac{4\left(\sum_{g\in J_3} \lambda_g^2\right)^{1/2}}{k_3 n^{1/2}} \right) ||X_{B^*}(\hat{V} - V^*)^T||_2.
\end{aligned}$$

Thus, we have

$$||X_{B^*}(\hat{V} - V^*)^T||_2^2 \leq 8\lambda_1 \left(\frac{s_1^{1/2}}{k_1 n^{1/2}} + \frac{s_2^{1/2}(\lambda 2/\lambda_1)\sigma_{D^+}}{k_2 n^{1/2}} + \frac{\left[\sum_{g\in J_3} (\lambda_g/\lambda_1)^2\right]^{1/2}}{k_3 n^{1/2}} \right) ||X_{B^*}(\hat{V} - V^*)^T||_2.$$

Plugging $\lambda_1 = 2Cd_X\sigma\sqrt{log(pm)}$ into above inequality and taking square on both sides, we have

$$\begin{aligned}
\frac{1}{n}||X_{B^*}(\hat{V} - V^*)^T||_2^2 &\leq \frac{64\lambda_1^2}{n} \left(\frac{s_1^{1/2}}{k_1 n^{1/2}} + \frac{s_2^{1/2}(\lambda 2/\lambda_1)\sigma_{D^+}}{k_2 n^{1/2}} + \frac{\left[\sum_{g\in J_3} (\lambda_g/\lambda_1)^2\right]^{1/2}}{k_3 n^{1/2}} \right) \\
&\leq 256 C^2 S^2 d_X^2 \sigma^2 \left[\frac{log(pm)}{n} \right].
\end{aligned}$$

Thus, we have

$$\begin{aligned}
||\hat{V} - V^*||_2^2 &\leq \frac{1}{n\sigma_{X_B^-}^2} ||X_{B^*}(\hat{V} - V^*)^T||_2^2 \\
&\leq \frac{256 C^2 S^2 d_X^2 \sigma^2}{\sigma_{X_B^-}^2} \left[\frac{log(pm)}{n} \right].
\end{aligned}$$

This completes proof of the theorem. □

B.2.7 Proof of Theorem 2

Proof. When conditions in Assumption 2 and Assumption 3 hold, we have

$$\begin{aligned}
||\hat{A} - A^*||_2 &= ||\hat{B}\hat{V}^T - B^*V^{*T}||_2 \\
&= ||\hat{B}\hat{V}^T - \hat{B}V^{*T} + \hat{B}V^{*T} - B^*V^{*T}||_2 \\
&\leq ||\hat{B}(\hat{V} - V^*)^T||_2 + ||(\hat{B} - B^*)V^{*T}||_2 \\
&\leq ||\hat{B}||_2||(\hat{V} - V^*)^T||_2 + ||(\hat{B} - B^*)||_2||V^{*T}||_2 \\
&\leq \left(||B^*||_2 + ||(\hat{B} - B^*)||_2\right) ||(\hat{V} - V^*)^T||_2 + ||(\hat{B} - B^*)||_2||V^{*T}||_2
\end{aligned}$$

Since we assume $||V^*||_2 = 1$, then $||B^*||_2 = ||B^*V^{*T}V^*||_2 = ||A^*V^*||_2 \leq \sigma_1(A) \leq \gamma d$. With fixed p and error bound of $\hat{V}$ obtained in Theorem 1, we have

$$\begin{aligned}
||\hat{A} - A^*||_2 &\leq \left(||B^*||_2 + ||(\hat{B} - B^*)||_2\right) ||(\hat{V} - V^*)^T||_2 + ||(\hat{B} - B^*)||_2||V^{*T}||_2 \\
&\leq rd||(\hat{V} - V^*)^T||_2 + ||(\hat{B} - B^*)||_2||(\hat{V} - V^*)^T||_2 + ||(\hat{B} - B^*)||_2 \\
&= ||(\hat{V} - V^*)||_2 + O(1/\sqrt{n}) \\
&= O_p\left[\left(\frac{\log(m)}{n}\right)^{1/2}\right]
\end{aligned}$$

Thus,

$$||\hat{A} - A^*||_2^2 = O_p\left[\frac{\log(m)}{n}\right]$$

Next, we have

$$\begin{aligned}
\frac{1}{n}||X(\hat{A} - A^*)||_2^2 &\leq ||X||_2^2||(\hat{A} - A^*)||_2^2 \\
&= O_p\left[\left(\frac{\log(m)}{n}\right)\right]
\end{aligned}$$

This completes proof of the theorem. □

Although we assume $||V||_2 = 1$ in Assumption 2, this is not limiting, as $||B^*||_2$ and $||V^*||_2$ are both bounded by $||A^*||_2$, which is restricted to be with finite singular values in a finite rank space defined in Assumption 2.

B.2.8 Derivation of ADMM solution

B.2.8.1 Update step In Section 3.3.2, we introduced definitions of K_ℓ, λ_ℓ, θ_ℓ, μ_ℓ and K. Let $\theta = (\theta_1, ..., \theta_N)^T$ be the concatenation of auxiliary variables $\theta_\ell = K_\ell V^v$ and let $\mu = (\mu_1, ..., \mu_N)^T$ be the concatenation of auxiliary variables μ_ℓ. We can write the augmented Lagrangian equation of (3.4) as:

$$\mathcal{L}_\rho(V^v, \theta, \mu) = \frac{1}{2}\|X_B^v V^v - Y^v\|_2^2 + \sum_{l=1}^{N} \lambda_\ell ||\theta_\ell||_2 + \sum_{l=1}^{N} \left[\mu_\ell^T(\theta_\ell - K_\ell V^v) + (\rho/2)||\theta_\ell - K_\ell V^v||_2^2\right]$$

Then, the iterative updates of V^v, θ_ℓ and μ_ℓ are

$$\begin{aligned}
V^{v(t+1)} &= \underset{V^v \in \Re^{(mr)}}{\arg\min}\, \mathcal{L}_\rho(V^v, \theta^{(t)}, \mu^{(t)}), \\
\theta_\ell^{(t+1)} &= \underset{V^v \in \Re^{(w_\ell)}}{\arg\min}\, \mathcal{L}_\rho(V^{v(t+1)}, \theta, \mu^{(t)}), \\
\mu_\ell^{(t+1)} &= \mu_\ell^{(t)} + \rho(\theta_\ell^{(t+1)} - K_\ell V^{v(t+1)}).
\end{aligned}$$

Thus, by solving differential equations $\frac{\partial \mathcal{L}_\rho(V^v, \theta^{(t)}, \mu^{(t)})}{\partial V^v} = 0$ and $\frac{\partial \mathcal{L}_\rho(V^{v(t+1)}, \theta, \mu^{(t)})}{\partial \theta} = 0$,

$$\begin{aligned}
V^{v(t+1)} &= \left(X^{vT}X^v + \rho K^T K\right)^{-1}\left[X^{vT} + K^T(\mu^{(t)} + \rho\theta^{(t)})\right], \\
\theta_\ell^{(t+1)} &= \left[1 - \lambda_\ell/(\rho||\eta_\ell^{(t)}||_2)\right]_+ \eta_\ell^{(t)}, \\
\mu_\ell^{(t+1)} &= \mu_\ell^{(t)} + \rho\left(\theta_\ell^{(t+1)} - K_\ell V^{v(t+1)}\right),
\end{aligned}$$

and updates of μ_ℓ depend on updates of V^v and θ_ℓ.

B.2.8.2 Stopping criteria In this algorithm, we use the stopping criteria described in Boyd et al. (2011). The algorithm will terminate when the primal and dual residuals converges to small values which achieve a linear combination of pre-specified levels of absolute (ϵ_{abs}) and relative (ϵ_{rel}) tolerance. Appropriate values for (ϵ_{abs}) and (ϵ_{rel}) depend on the specific application and scale of the data. Let the primal and dual residuals at iteration time t be $r^{(t)} = \theta^T - KV^{v(t)}$ and $s^{(t)} = \rho K^T(\theta^{(t)} - \theta^{(t-1)})$, respectively. Let $|\theta^{(t)}|$ represents the number of elements in $\theta^{(t)}$. The stopping criteria are $||r^{(t)}||_2 \leq \epsilon_{pri}^{(t)}$ and $||s^{(t)}||_2 \leq \epsilon_{dual}^{(t)}$, where $\epsilon_{pri}^{(t)}$ and $\epsilon_{dual}^{(t)}$ are primal residual dual residual tolerance at iteration time t respectively:

$$\epsilon_{pri}^{(t)} = \sqrt{\rho}\epsilon_{abs} + \epsilon_{rel}\max\left(||KV^{v(t)}||_2, ||\theta^{(t)}||_2\right),$$
$$\epsilon_{dual}^{(t)} = \sqrt{|\theta^{(t)}|}\epsilon_{abs} + \epsilon_{rel}||K^T\mu^{(t)}||_2,$$

Bibliography

Admiraal-Behloul, F., Van Den Heuvel, D., Olofsen, H., van Osch, M. J., van der Grond, J., Van Buchem, M., and Reiber, J. (2005). Fully automatic segmentation of white matter hyperintensities in mr images of the elderly. Neuroimage **28,** 607–617.

Anbeek, P., Vincken, K. L., Van Osch, M. J., Bisschops, R. H., and Van Der Grond, J. (2004). Probabilistic segmentation of white matter lesions in mr imaging. NeuroImage **21,** 1037–1044.

Anitha, M., Selvy, P. T., and Palanisamy, V. (2012). Wml detection of brain images using fuzzy and possibilistic approach in feature space. WSEAS Transactions on computers, e-ISSN **22242872,**.

Beer, J. C., Aizenstein, H. J., Anderson, S. J., and Krafty, R. T. (2019). Incorporating prior information with fused sparse group lasso: Application to prediction of clinical measures from neuroimages. Biometrics .

Bigos, K. L. and Weinberger, D. R. (2010). Imaging genetics—days of future past. Neuroimage **53,** 804–809.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning **3,** 1–122.

Bunea, F., She, Y., Wegkamp, M. H., et al. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. The Annals of Statistics **39,** 1282–1309.

Caligiuri, M. E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., and Cherubini, A. (2015). Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review. Neuroinformatics **13,** 261–276.

Chen, K., Chan, K.-S., and Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **74,** 203–221.

Chen, K., Dong, H., and Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. Biometrika **100,** 901–920.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. Journal of the American Statistical Association **107,** 1533–1545.

Cohen, A. D., Mowrey, W., Weissfeld, L. A., Aizenstein, H. J., McDade, E., Mountz, J. M., Nebes, R. D., Saxton, J. A., Snitz, B., DeKosky, S., et al. (2013). Classification of amyloid-positivity in controls: comparison of visual read and quantitative approaches. Neuroimage **71,** 207–215.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In Proceedings of the 23rd international conference on Machine learning, pages 233–240. ACM.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. Ecology **26,** 297–302.

Dyrby, T. B., Rostrup, E., Baaré, W. F., van Straaten, E. C., Barkhof, F., Vrenken, H., Ropele, S., Schmidt, R., Erkinjuntti, T., Wahlund, L.-O., et al. (2008). Segmentation of age-related white matter changes in a clinical multi-center study. Neuroimage **41,** 335–345.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. Journal of Statistical Software **40,** 1–18.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33,** 1–22.

Ghafoorian, M., Karssemeijer, N., Heskes, T., Uden, I. W., Sanchez, C. I., Litjens, G., Leeuw, F.-E., Ginneken, B., Marchiori, E., and Platel, B. (2017). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Scientific Reports **7,** 5110.

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In European Conference on Information Retrieval, pages 345–359. Springer.

Haralick, R. M., Sternberg, S. R., and Zhuang, X. (1987). Image analysis using mathematical morphology. IEEE transactions on pattern analysis and machine intelligence pages 532–550.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

Herskovits, E., Bryan, R., and Yang, F. (2008). Automated bayesian segmentation of microvascular white-matter lesions in the accord-mind study. Advances in medical sciences **53,** 182–190.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12,** 55–67.

Horn, R. A. and Johnson, C. R. (1991). Topics in Matrix Analysis. Cambridge University Press.

Howard, A., Sandler, M., Chu, G., Chen, L., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. (2019). Searching for mobilenetv3. arxiv 2019. arXiv preprint arXiv:1905.02244 .

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 .

Hsu, D. C., Mormino, E. C., Schultz, A. P., Amariglio, R. E., Donovan, N. J., Rentz, D. M., Johnson, K. A., Sperling, R. A., and Marshall, G. A. (2016). Lower late-life body-mass index is associated with higher cortical amyloid burden in clinically normal elderly. Journal of Alzheimer's Discovery **53,** 1097–1105.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Itti, L., Chang, L., and Ernst, T. (2001). Segmentation of progressive multifocal leukoencephalopathy lesions in fluid-attenuated inversion recovery magnetic resonance imaging. Journal of Neuroimaging **11,** 412–417.

Karim, H., Tudorascu, D., Cohen, A., Price, J., Lopresti, B., Mathis, C., Klunk, W., Snitz, B., and Aizenstein, H. (2019). Relationships between executive control circuit activity, amyloid burden, and education in cognitively healthy older adults. Am J Geriatr Psychiatry **12,** 1360–1371.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kossaifi, J., Bulat, A., Tzimiropoulos, G., and Pantic, M. (2019). T-net: Parametrizing fully convolutional nets with a single high-order tensor. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7822–7831.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105.

Lao, Z., Shen, D., Liu, D., Jawad, A. F., Melhem, E. R., Launer, L. J., Bryan, R. N., and Davatzikos, C. (2008). Computer-assisted segmentation of white matter lesions in 3d mr images using support vector machine. Academic radiology **15,** 300–313.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE **86,** 2278–2324.

Lee, D., Kwon, S. J., Kim, B., and Wei, G.-Y. (2019). Learning low-rank approximation for cnns. arXiv preprint arXiv:1905.10145 .

Ma, Z. and Sun, T. (2014). Adaptive sparse reduced-rank regression. arXiv preprint arXiv:1403.1922 .

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics **19,** 1236–1246.

Muschelli, J., Gherman, A., Fortin, J. P., Avants, B., Whitcher, B., Clayden, J. D., Caffo, B. S., and Crainiceanu, C. M. (2018). Neuroconductor: an R platform for medical imaging analysis. Biostatistics .

Muschelli, J., Sweeney, E., Lindquist, M., and Crainiceanu, C. (2015). fslr: Connecting the fsl software with r. The R journal **7,** 163.

Nadkarni, N., D, T., E, C., BE, S., AD, C., Halligan E, a. M. C., HJ, A., and WE, K. (2019). Association between amyloid-, small-vessel disease, and neurodegeneration biomarker positivity, and progression to mild cognitive impairment in cognitively normal individuals. J Gerontol A Biol Sci Med Sci. **11,** 1753–1760.

Nadkarni, N. K., Tudorascu, D., Campbell, E., Snitz, B. E., Cohen, A. D., Halligan, E., Mathis, C. A., Aizenstein, H. J., and Klunk, W. E. (2019). Association between amyloid-β, small-vessel disease, and neurodegeneration biomarker positivity, and progression to mild cognitive impairment in cognitively normal individuals. The Journals of Gerontology: Series A .

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458 .

Pillai, J. J., Friedman, L., Stuve, T. A., Trinidad, S., Jesberger, J. A., Lewin, J. S., Findling, R. L., Swales, T. P., and Schulz, S. C. (2002). Increased presence of white matter hyperintensities in adolescent patients with bipolar disorder. Psychiatry Research: Neuroimaging **114,** 51–56.

Reinsel, G. C. and Velu, R. P. (1998). Multivariate reduced-rank regression. Springer, New York.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4510–4520.

Schmidt, P. (2017). Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. PhD thesis, lmu.

She, Y. and Chen, K. (2017). Robust reduced-rank regression. Biometrika **104,** 633–647.

Sheline, Y. I., Price, J. L., Vaishnavi, S. N., Mintun, M. A., Barch, D. M., Epstein, A. A., Wilkins, C. H., Snyder, A. Z., Couture, L., Schechtman, K., et al. (2008). Regional white matter hyperintensity burden in automated segmentation distinguishes late-life depressed subjects from comparison subjects matched for vascular risk factors. American Journal of Psychiatry **165,** 524–532.

Shinohara, R. T., Goldsmith, J., Mateen, F., Crainiceanu, C., and Reich, D. (2012). Predicting breakdown of the blood-brain barrier in multiple sclerosis without contrast agents. American Journal of Neuroradiology **33,** 1586–1590.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data **6,** 60.

Simões, R., Mönninghoff, C., Dlugaj, M., Weimar, C., Wanke, I., van Walsum, A.-M. v. C., and Slump, C. (2013). Automatic segmentation of cerebral white matter hyperintensities using only 3d flair images. Magnetic resonance imaging **31,** 1182–1189.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Smith, S. M. (2002). Fast robust automated brain extraction. Human brain mapping **17,** 143–155.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15,** 1929–1958.

Sweeney, E. M., Shinohara, R. T., Shiee, N., Mateen, F. J., Chudgar, A. A., Cuzzocreo, J. L., Calabresi, P. A., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. (2013). Oasis is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in mri. NeuroImage: clinical **2,** 402–413.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9.

Team, R. C. et al. (2013). R: A language and environment for statistical computing.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58,** 267–288.

Tibshirani, R. J. (2011). The solution path of the generalized lasso. Stanford University.

Tublin, J. M., Adelstein, J. M., del Monte, F., Combs, C. K., and Wold, L. E. (2019). Getting to the heart of alzheimer disease. Circulation Research **124,** 142–149.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4itk: improved n3 bias correction. IEEE transactions on medical imaging **29,** 1310–1320.

Valcarcel, A. M., Linn, K. A., Vandekar, S. N., Satterthwaite, T. D., Muschelli, J., Calabresi, P. A., Pham, D. L., Martin, M. L., and Shinohara, R. T. (2018). Mimosa: an automated method for intermodal segmentation analysis of multiple sclerosis brain lesions. Journal of Neuroimaging **28,** 389–398.

Van Den Heuvel, D., Admiraal-Behloul, F., Ten Dam, V., Olofsen, H., Bollen, E., Murray, H., Blauw, G., Westendorp, R., De Craen, A., Van Buchem, M., et al. (2004). Different progression rates for deep white matter hyperintensities in elderly men and women. Neurology **63,** 1699–1701.

Wen, W., Wu, C., Wang, Y., Chen, Y., and Li, H. (2016). Learning structured sparsity in deep neural networks. In Advances in neural information processing systems, pages 2074–2082.

West, N. A. and Haan, M. N. (2009). Body adiposity in late life and risk of dementia or cognitive impairment in a longitudinal community-based study. Journal of Gerontology **64,** 103–109.

Wong, T. Y., Klein, R., Sharrett, A. R., Couper, D. J., Klein, B. E., Liao, D.-P., Hubbard, L. D., Mosley, T. H., investigators, A., et al. (2002). Cerebral white matter lesions, retinopathy, and incident clinical stroke. Jama **288,** 67–74.

Wu, M., Rosano, C., Butters, M., Whyte, E., Nable, M., Crooks, R., Meltzer, C. C., Reynolds, C. F., and Aizenstein, H. J. (2006). A fully automated method for quantifying and localizing white matter hyperintensities on mr images. Psychiatry Research: Neuroimaging **148,** 133–142.

Yoo, B. I., Lee, J. J., Han, J. W., Lee, E. Y., MacFall, J. R., Payne, M. E., Kim, T. H., Kim, J. H., Kim, K. W., et al. (2014). Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images. Neuroradiology **56,** 265–281.

Yu, X., Liu, T., Wang, X., and Tao, D. (2017). On compressing deep models by low rank and sparse decomposition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7370–7379.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B **68,** 49–67.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE transactions on medical imaging **20,** 45–57.